DISSERTATION SYNOPSIS

On

# 'Sentiment Analysis of transliterated hindi and marathi script'

Submitted in partial fulfillment of the requirements for the degree of

**Master of Engineering** in **Information Technology (AI and Robotics)**

By

**Mr. Mohammed Arshad Ansari**

Under the guidance

of

**Prof. Sharvari Govilkar**

(Asst. Prof. Computer Department)


**DEPARTMENT OF INFORMATION TECHNOLOGY**
**PILLAI INSTITUTE OF INFORMATION TECHNOLOGY, ENGINEERING, MEDIA STUDIES & RESEARCH**
**NEW PANVEL - 410206**
**UNIVERSITY OF MUMBAI**
**Academic Year 2015-16**

# DEPARTMENT OF INFORMATION TECHNOLOGY
# PILLAI INSTITUTE OF INFORMATION TECHNOLOGY, ENGINEERING, MEDIA STUDIES & RESEARCH
# NEW PANVEL - 410206
# UNIVERSITY OF MUMBAI
# Academic Year 2015-16


## SYNOPSIS OF PROJECT WORK

| | |
|---|---|
| Name of the Dissertation: | Sentiment Analysis of transliterated hindi and marathi script |
| Student's Name: | Mr. Mohammed Arshad Ansari |
| Class: | M.E. (Information Technology) |
| College: | PIllai Institude of Information Technology, New Panvel |
| Semester: | IV |
| University Registration Number: | |
| Date of Registration: | |
| Exam Fee Receipt No.: | 12698 |
| Name of Guide: | Prof. Sharvari Govilkar |

| Semester | Exam Seat Number | Result (SGPI) |
|---|---|---|
| 1st | 4651 | 7.00 |
| 2nd | 4501 | 6.14 |

# Abstract

There is a growing research on sentiment analysis of various languages, which is being supplanted heavily by those same techniques and methods being applied on the mix code or transliterated text for the same purpose. This growing research is a result of necessity created through the advent of social media as well as textual analysis of the data being collected online. This paper, rather than being a pioneer, is about extending that research for further improvement. Herein, we assess the existing status, standards and achievements of the researchers in the given field and supplant it with out proposed methology to increase precision.

Although, the current work is a proposal with improvements over established techniques, it is also however gonig to be quite comparative when it comes to the existing findings. The idea is to not just improve what has already been built or shown to be true, but also check if the simplest approach is still the best way to proceed or not. By this we mean the existing direct supervised learning for sentiment analysis, without much NLP or language specific work.

Since we shall be testing our approach against the existing state of the art as well as entering the area previously not under coverge (marathi transliterated text), this work is bound to make great strides in the field of sentiment analysis.

# 1 Introduction

Sentiment analysis is a process of analysing natural language and figuring out the sentiments involved or expressed through the source material, with respect to the topic. The basic idea behind sentiment analysis is that each textual sentence may or may not contain some kind of polarity, expressing a degree of emotions along with the information. It is much easier to read in to those polarity when the text is spoken and not written due to the tone of the speaker; whereas, in case of written text, it is the context that is useful while determining the polarities in the statements. Sentiment analysis has grown to be one of the most important research areas

when it comes to textual analysis on the web. Reason being, obviously, is to be able to make sense of the data as well as to understand the tone of information being provided. There are numerous application, ranging from product/customer support review to improve quality of service (QOS) by corporations to understanding geo-political motivations when certain news breaks. People react on social media, especially when they are charged emotionally and when emotions take the form of textual content to vent, it has been observed that it does in a manner which is more close to a person's mother tongue.

Hindi is spoken by more than 500 million people around the world, making it one of the most spoken language in the world. Besides, English has turned out to be an international language, a lot of people speak English on the internet, however; as described above, there are instances when people use english language to phonetize and express in a foreign language. This is seen far more in India subcontinent, where people prefer to write using english alphabets, but most often, use the words from the mother tongue. If we only look at all the youtube comments (especially if they are about some controversial issues), we would see a lot of usage of such transliterated messages or mix-script writing. Another behavior worth noting is related to vocabulary. People from subcontinent use words such as 'Bye', 'Thank you', 'Good night', 'Please', 'Sorry' and intermix them with their native tongues. This mixture of language has been observed profoundly at varying levels of society. Therefore, it would not be very far-fetched to say that the languages are evolving by mixing language themselves. This forms the necessary reason for why there needs to be analysis of mixed-languages and it starts with analyzing that which is mostly available, the mixed-script. Here, we are not going to invent something new, nor are we going to do something entirely differently. However, the purpose behind this work is to stand on the shoulders of giants and take the research of what has already been done to what it can be. This we strive to do, by improving the performances by innovatively applying techniques which have worked better in other cases. Therefore, as it will be seen, our proposed approach as well is a mixture of disparate attempts in varying domains (even slightly) to come together for better whole.

Sentiment analysis is a lot tougher for languages that are outside eurozone, due to their lexical syntax being very different from european languages as well as due to majorly, less

amount of work being done on it. Semantic analysis requires annotated text corpus to train classifers, which is most of the time a very huge manual task. It has been undertaken for English and for many other European langauges, while at the same time, work from one supplementing work for another language, due to the similarities existent in those languages. When it comes to languages such as Hindi and Marathi, such resources are very less compared to the above mentioned languages. Moreso for Marathi, since a lot of work has been done and progress made in case of Hindi. Most of the reason for the under development of the research for these languages are (1) Not much annotated textual corpus needed for traning, (2) Lack of basic language tools like taggers and parsers. These problems will be solved in time and this work is a part of all the woks which will finally solve this problem. Having expressed the problem, in this work, we also laude the work that has already gone in to this respective field, without which this would not have been possible. It is really interesting to note, that a great amount of effort has just started pouring in for this particular part of sentiment analysis. It is naught with great anticipation that this work is being progressed. Besides, as the sentiment analysis of the textual data being to shape more and more, the greater the benefit will be to the field of general AI. When it comes to human capacity, not representing emotions would be the biggest gap in the domain, which exactly is sentiment analysis has started to fill.

# 2 Literature Review

Code mixing has been done for more than a couple decades and was investigated during initial period by Gold [8] for the purpose of language identification. The same phenomenon for Indian langauges was worked upon by Annamalai [1], pioneering the research field for the subcontinent languages. Recently, it was investigated by Elfarti et. al. [7] and was termed as linguistic code switching by the research group. Karimi [11] made the case for machine translation for the purpose of transliteration in the survey and suggested transliteration based on phoneme based approach and transliteration generation using bilingual corpus, while presenting the key issues that arise during the transliteration process. Dewaele [6] pointed out the strong emotional presence as being the main marker for the existence of code switch that happens in textual corpus. Gupta et. al. [9] mined the tranlisteration pairs between hindi and english from the music lyrics of bollywood songs for Fire'14 shared task, which is quite handy

for training in language sentiments.

The issue of identification of language of the code - mix script is another challenge that has been answered by the research community. A statistical approach was proposed by Kundu and Chandra et. al. [14] for the automatic detection of English words in Bengali + English (Benglish) text. A conditional random field model for weakly supervised learning model was used for word/token labelling by King and Abney [13] with a good result of > 90%. Barman [5] used facebook user data for identifying the language in mixed script and concluded that the supervised learning outperforms the dictionary based approaches. POS Tagging and transliteration efforts for Hindi + English data on social media was experimented upon by Vyas et. al. [21] and came to the conclusion that any operation on transliteration text will largely benefit from pos tagging.

Although sentiment analysis is being worked upon for quite some time now and it has already entered the mainstream application. There are works being done transliteration of Indian languages, out of which some have been listed below under this section.

Joshi et.al. [10] performed experiment to compare three approaches for the sentiment analysis of hindi text and found that HSWN performs better than Machine Translation approach but under performs in language traning of sentiment corpus in hindi. This was, however; performed in 2010 and the HSWN has been continually improving past these experiments. The same result was reiterated with by Balamurali et. al. [2]. Kashyap [12] found a way to perform Hindi Word Sense Disambiguation using wordnet with encouraging results for nouns. Subjective lexical resource was developed by Bakliwal et. al. [3] by using only wordnet and graph traversal algorithm for adverbs and adjectives.

Balamurali A R et. al. [4] performed experiement to figure out in language supervised training of sentiments against the machine translated source for sentiment analysis. They found that the MT based approach under performs much worse compared to in language traning of sentiment. Fuzzy Logic membership function was used to determine the degree of polarity of the sentiment for a given POS tagged preposition by Rana [20]. Hindi Senti WordNet was developed by Balamurali A. R. et. al. [4] using the Senti Wordnet by using linked wordnet. HSWN along with negetion discourse was applied by Pandey [17] and Mittal et. al.

[16] or sentiment analysis of Hindi language text corpora, with the accuracy of 80.21 achived.

There is only one work done on the sentiment analysis of hindi transliteration by Srinivas ([18] and [19] ) and the approach taken was to tag words with identified language and then run against respective POS tagger for languages and sentiment analysis done on the output. The approach yielded 85% percision. Although not much has been done on marathi front, however; it is still in the works and pipeline.

# 3 Problem Statement

*The proposed work is geared towards the problem of sentiment analysis of code - mix script (Romanized) for languages such as hindi and marathi. The current literature has brought the accuracy up to 85 percent for hindi transliterated text, which this work aims to take to 90-95 percent accuracy and paritially accomplish for marathi language as well.*

# 4 Methodology

## 4.1 Scope of work

The current word considers hindi/marathi text in romanized script as input which may contain phonetic words, sounds; however, it is not considering social language like gr8, rt, f9, etc as input source. For now it is considered as noise for the result of this work. Although, we are not performing sentiment analysis on English text as part of this text, which is simply because it has been under taken in many works preceeding this one. Therefore, concentration will solely be on the text which is transliterated hindi or marathi. Also, the architecture of proposed approach has intergration in mind and hence, it will be able to plug social sentiment analysis or twitter sentiment analysis or plain english analysis and will work only for the transliterated text, while taking inputs from the mentioned analyzers for their established polarity. The results can be merged and shown to have improved the overall accuracy.

## 4.2 Proposed Approach

There are going to me multiple approach for testing to be implemented as part of this work. The purpose of those work will be to ensure that the proposed system performs better than what has been accomplished by other researchers. Although, here we will only go into the actual proposed system to understand its working and predict the possible improvements.
The proposed approach we are to take in this work comprises of extending the work of Srinivas [19] with multiple improvement points at multiple level of the process. Each step is listed below with the improvement suggested from this work. All the work done on that paper has been uploaded to the website [15], which we shall be using in this paper extensively and building on top of it.

### 4.2.1 Text Normalization

Text normaliation step depends heavily on work of Srinivas ([18]), which has following steps:

- ***Language Identification***; Tagging of the words as <word>/<tag>, where <tag> can be E or english, H for hindi and M or marathi,

- ***Spelling Corrections***; There are multipe ways to write Mujhe in hindi such as muze, muje, etc. To come to common and widely used spelling becomes very importnat

- ***Ambiguous words***; Words such as 'me' means same thing in english and marathi, where as in hindi it is sometimes used to say inside with another spelling being 'mein'.

- ***Sounds***; Words such as aww, oohh, ouch, ewww, etc. They do contain rich information when it comes to sentiments.

- ***Phonetic words***; Words such as pleej usually is misspelling of word please, spoken in some areas of the subcontinent. It gets written too in the similar manner.

- ***Transliteration***; Conversion of hindi/marathi words written in english to appropriate devnagari script.

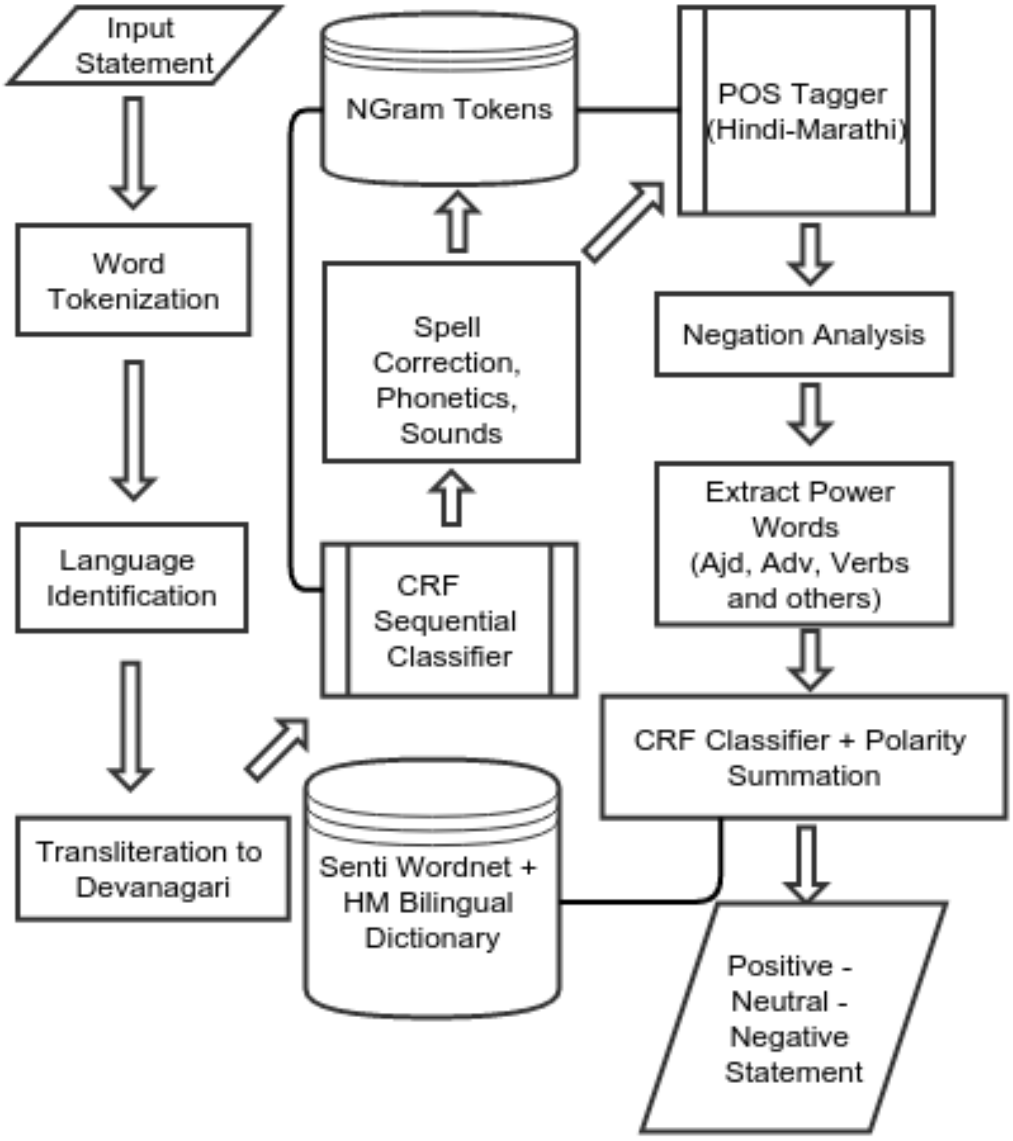


Figure1: Flow diagram for the proposed approach

All then above enumerated steps have been covered by Srinivas ([18]) and doesn't require us to go in the details of those, however; we will look at some of those steps to get a proper grip on the subject. At this point, this work will simply reuse those steps for hindi and try to closely perform them for marathi as well.

### 4.2.2 POS Tagging, Discourse analysis, Senti Word Net

Before sentiment analysis can be performed, it is necessary to deal with few important things. We are striving to extend and improve upon earlier work such as Srinivas ([18]) and therefore following much in the same footsteps. Both the step are explained further below. Once we have the document in english or hindi, the next step is to run it through POS tagger based on respective language. The approach will be straight forward as detailed here [21]. The POS Tagged prepositions then shall be run through the negation discourse analysis to invert the POS tagged adjectives and adverbs in case of negetive discourse as explained by Pandey [17] and Mittal et. al. [16]. The output of POS Tagger shall be used to look up sentiword identifier for the word groups using Sentiwordnet or HSWN, for English and Hindi, respectively. HSWN has been improved by Pandey [17] by making additions to it and that will be used in this work. Here, there are three major improvements we are considering. Since, it was established [19] that the basis for sentiment analysis being POS tagged adjectives and adverbs gives much better result that depending direcly on lexicon or wordnet look up for each work, we would be going that route. Secondly, addition of discourse analysis would further enhance on the existing work [19].

Once we get sentiword identifier for each token - word, next step is to put it through the classifier which will give the polartiy of the statement provided. This polarity checking decision can be as simple as simple summation of all word-token sentiment polarities or further analysis can be performed to figure out what really is the polarity of word-token and its membership with negetive, postive or neutral . This step being the vital one can be accomplished using most trust classifier like SVM, Random Forests, however; impetus shall be given on naive bayes classifier for brevity's sake.

Quite clearly, we have input as POS tagged statements with greater emphasis on adjectives and adverbs that are inverted in case of negation present in the preposition. Once we have this tagged information, we would like to test on both the process of simple polarity count summation of the given input and traning the classifier, in order to come up with the best possible result in terms of accuracy. As can be seen in the figure, the classification uses Senti Wordnet for polarity database, which is combination of HSWN and Hindi-Marathi Bilingual Dictionary to match sentiments across language and to look up is polarity.

# 5 Requirement Analysis

- ***Hardware requirements***;
  * i5 Processor
  * 8 GB Ram
  * 200 GB + HDD Space
  * Ubuntu OS 15.10

- ***Software requiremetns and Tools***;
  * Python 2.7
  * NLTK
  * Scikit Learn
  * Numpy

- ***Data Requirment***;
  * Code - Mix Data from Fire @ 2013 and 2014
  * Linugistic weebly data [18]

# 6 Conclusion

The most important aspect of this work i.e. the results are what is coming next. We will show that the approach proposed in this work performs better than all the work presented here in literature, when considered independently. It is the synergy, which the approach presented there, promises. The implementation will happen for all ways that differ from the approach too, so that comparisons can be made and conclusions drawn without the strawman arguments.

There is a lot of work to be performed before any concrete conclusion can be expressed, however; There is a great possibility that the approach suggested in the given work will result in improvement in the field of sentiment analysis, that can again be extended for greater

language coverage in as well as out of indian languages. These strides towards such improvements will result in machine's being able to understand human sentiments better, which is one of the greatest challenge being faced by the research in general AI. Ours is but a small step towards that goal. It will not be too far fetched to believe that the improvements will range from 5 to 10 percent improvement where we will see the accuracy reach 95 percent.

# References


[1] E. Annamalai. "The anglicized Indian languages: A case of code mixing". In: *International Journal of Dravidian Linguistics* 7.2 (1978), pp. 239–47.

[2] Balamurali A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya. "Lost in Translation: Viability of Machine Translation for Cross Language Sentiment Analysis". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by David Hutchison et al. Vol.7817. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 38–49. ISBN: 978-3-642-37255-1 978-3-642-37256-8. URL: http://link.springer.com/10.1007/978-3-642-37256-8_4 (visited on 11/11/2015).

[3] Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. "Hindi subjective lexicon: A lexical resource for hindi polarity classification". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*. 2012. URL: http://web2py.iiit.ac.in/research_centres/publications/download/inproceedings.pdf.a92646aa66336f21.4c5245432731322d3637332e706466.pdf (visited on 12/07/2015).

[4] Balamurali A R, Aditya Joshi, and Pushpak Bhattacharyya. "Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets". In: DOI: 10.1.1.379.149.

[5] Utsab Barman et al. "Code Mixing: A Challenge for Language Identification in the Language of Social Media". In: *EMNLP 2014* (2014), p. 13.

[6] Jean-Marc Dewaele. "Emotions in multiple languages". In: (2010).

[7] Heba Elfardy and Mona T. Diab. "Token Level Identification of Linguistic Code Switching." In: *COLING (Posters)*. 2012, pp. 287–296.

[8] E. Mark Gold. "Language identification in the limit". In: *Information and control* 10.5 (1967), pp. 447–474.

[9] Kanika Gupta, Monojit Choudhury, and Kalika Bali. "Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics." In: *LREC*. 2012, pp. 2459–2465. URL: http://lrec.elra.info/proceedings/lrec2012/pdf/365_Paper.pdf (visited on 11/11/2015).

[10] Aditya Joshi, A. R. Balamurali, and Pushpak Bhattacharyya. "A fall-back strategy for sentiment analysis in hindi: a case study". In: *Proceedings of the 8th ICON* (2010). URL: http://www.cse.iitb.ac.in/~balamurali/papers/ICON%20229.pdf (visited on 11/11/2015).

[11] Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. "Machine transliteration survey". In: *ACM Computing Surveys (CSUR)* 43.3 (2011), p. 17.

[12] Prabhakar Pandey Laxmi Kashyap. "Hindi Word Sense Disambiguation". In: (). URL: http://megha.garudaindia.in/iitb-nlp/hindiwn/papers/HindiWSD.pdf (visited on 12/07/2015).

[13] Ben King and Steven P. Abney. "Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods." In: *HLT-NAACL*. 2013, pp. 1110–1119.

[14] Bijoy Kundu and Swarup Chandra. "Automatic detection of English words in Benglish text: A statistical approach". In: *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on*. IEEE, 2012, pp. 1–4. ISBN: 1-4673-4367-6.

[15] *Linguistic resources website, Sharma, Shashank and PYKL Srinivas and Balabantaray, Rakesh Chandra work on Text normalization of code mix and sentiment analysis*. URL: http://linguisticresources.weebly.com/downloads.html (visited on 12/12/2015).

[16] Namita Mittal et al. "Sentiment Analysis of Hindi Review based on Negation and Discourse Relation". In: *Sixth International Joint Conference on Natural Language Processing*. 2013, p. 45. URL: http://www.basantagarwal.com/wp-content/uploads/2013/02/IJCNLP-WORKSHOP_W13-4306.pdf (visited on 12/07/2015).

[17] Pooja Pandey and Sharvari Govilkar. "A Framework for Sentiment Analysis in Hindi using HSWN". In: *International Journal of Computer Applications* 119.19 (2015). URL: http://search.proquest.com/openview/1549cf02e0848335d2dd7a268e05f025/1?pq-origsite=gscholar (visited on 11/11/2015).

[18] Shashank Sharma, PYKL Srinivas, and Rakesh Chandra Balabantaray. "Text normalization of code mix and sentiment analysis". In: *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*. IEEE, 2015, pp. 1468–1473. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7275819 (visited on 11/13/2015).

[19] Shashank Sharma and PYKL Srinivas. *Sentiment Analysis of Code - Mix Script*.

[20] Shweta Rana. "Sentiment Analysis for Hindi Text using Fuzzy Logic". In: (Aug. 2014). ISSN: 2249-555X. URL: http://www.worldwidejournals.com/ijar/file.php?val=August_2014_1406901800__111.pdf (visited on 12/11/2015).

[21] Yogarshi Vyas et al. "Pos tagging of english-hindi code-mixed social media content". In: *Proceedings of the First Workshop on Codeswitching, EMNLP*. 2014.