ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

Importance and Challenges of Social Media Text

Shailendra Kumar Singh
Research Scholar, CSE Department
Sant Longowal Institute of Engineering & Technology
Sangrur, India

Manoj Kumar Sachan
Associate Professor, CSE Department
Sant Longowal Institute of Engineering & Technology
Sangrur, India

Abstract: The rapid growth of social media like twitter, Facebook, WhatsApp, messenger etc. has increased the availability of unstructured data (texts, images, videos) amount on internet. These texts are different from traditional text. The text written on social media are called Social Media Text. In India more than 50% comments on social media are written in Indian Languages using Unicode and Phonetic typing. The Pre-processing of these texts for application of Natural Language Processing (NLP) is a challenging task. This paper will help the researchers to understand the concept of code mixing, social media text, code mixed text and various challenges of social media text.

Keywords: social media; social media text; code mixing, POS tagger, Language Identification.

I. INTRODUCTION

Human being is the most intelligent living thing in the world. The best God given gift for human being is the ability to communicate with others. With the growth in science and technology the way and style of communication has been changed. There are various medium of communication like spoken form, written form and sign language etc. With time different kind of effective languages has been developed for smooth communication. India is a country of multilingual societies where people speaks more than one language and also mix multiple languages during communication. In the past mixing of multiple languages was allowed only in spoken form, but the development in social media has allowed people to mix multiple languages for effective communication in written form. So the code mixing (language mixing) is a new concept in the field of NLP applications.

Social networking websites play a vital role in user's online activities. India is set to have the highest internet protocol traffic growth with a 44% (compound annual growth rate) between 2012 and 2017, followed by Indonesia (42%) and South Africa (31%) [1]. The Worldwide Internet users have increased rapidly as shown in table-1[2]. Ten year back, social sites like Skype, facebook, YouTube, twitter and instagram etc. did not exist. Day by day the amount of user generated content has exponentially increased on these social networking platforms [3]. These social media contain text written in single, multiple languages or even in phonetic typing. However these social media texts contain essential information about products, issues, service etc. The rest part of this paper focuses on social media text (SMT), types of text and challenges of SMT during pre-processing.

Table I. WORLD WIDE INTERNET USERS

Year	*World Population*
2005	6.5 Billion
2010	6.9 Billion
2014	7.2 Billion

II. RELATED WORK

Part-of-Speech (POS) tagging is one of the pre-processing techniques for NLP applications. POS tagger is widely used in Sentiment Analysis (SA) based on subjective lexicon approach. POS tagging for monolingual text has been studied extensively with an accuracy as high as 97.3% for some languages (K. Toutanova et al., 2015 [4]). S.K Singh et al., 2015 [5] used POS tagger to identify "verb" in a sentence for SA. But, still very less work has been done on POS tagging of Social Media Code Mixed Text (SMCMT). Y. Vyas et al., 2014 [6] developed POS tagger for English-Hindi Code Mixed text (CMT). They have used 12 POS tags which are applicable to both English and Hindi languages. Also R. Sequiera et al., 2015 [7] have developed POS tagger for Hindi-English Code Mixed Text from Social Media. R.N.Patel et al., 2016 [8] have proposed POS tagger for Hindi-English, Bengali-English and Telugu-English code mixed data using Recurrent Neural Network Language Model (RNN-LM) architecture. S. Ghosh et al.,2016 [9] have presented POS tagger for Hindi-English, Bengali-English and Tamil-English using Conditional Random Field and a sequence learning method with an accuracy 73.2%, 75.22% and 64.83 % respectively.

Dey and fung, 2014 [10] presented the collection of a Hindi-English code switching corpus, which includes total nine student interviews. The students are proficient in both Hindi and English. On an average, roughly 67% of each sentence was made up of Hindi words and 33% English words.

Barman et al., 2014 [11] presented a study on social media communication of Indian Language code mixing for automatic language identification. They performed word level language identification experiment on Code mixing dataset between Bengali, English and Hindi. They used different techniques such as dictionary-based approach (93.64% accuracy), SVM (with or without context) and conditional Random Field (CRF) with 95.76% accuracy.

S. Dutta et al., 2015 [12] presented techniques to solve the problem of spelling error and language identification for code mixed social media text (English-Bengali). They have used Conditional Random Field (CRF) and post processing heuristics approaches to develop the word level language identifier with 90.5% accuracy. They have developed Spell Checker for text written in English language using Noisy Channel model with 69.43%. They have solved the problem of

wordplay, contracted words and phonetic spellings using language model.

III. SOCIAL MEDIA TEXT

The culture and society influence the spoken or written language. In India people stay together even if they belong to different caste, culture, state, religion etc. They write or speak in different languages, but their pronunciation changes according to place, society, culture etc [13]. In India people speak more than 20 languages. The English is an universal spoken language, which connects people all over the world. In this world a large proportion of people are bilingual. The communication in Bilingual is a very common phenomenon. The interest of people in Bilingual increases the chance to mix two languages. In this way the concept of writing in bilingual or code mixing arises. With the rapid growth of modern technologies and internet facilities increases the importance of social media in human life. Various types of electronic devices and software support to write text in bilingual on social media.

The text written in news paper, letters, articles or anywhere for formal communication are called Formal Text, while text used on social media are called Social Media Text (SMT). SMT can categories into SMT written in monolingual and SMT written in Bilingual or Multilingual as shown in Fig.1.

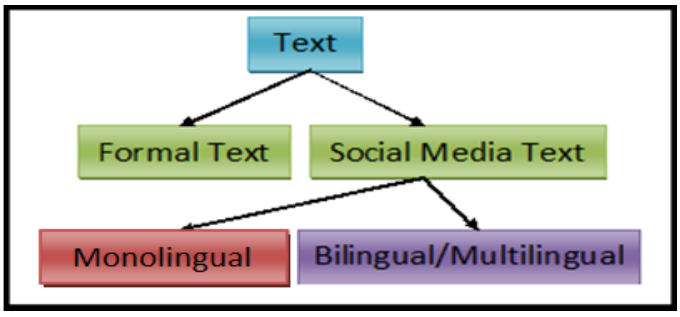

Figure 1 Types of Text

A. *SMT Written in Monolingual*

In this case, the texts are written only in single language using Unicode or Phonetic Typing (using roman characters). In Fig.2, text is written in Hindi language using Devanagari Script while in Fig.7 written in Hindi language using Phonetic typing. In India People writes 33.54% comments using phonetic typing and only 17.94% using Unicode in monolingual (Indian Language) as shown in table 2, based on (total 477) social media comments.

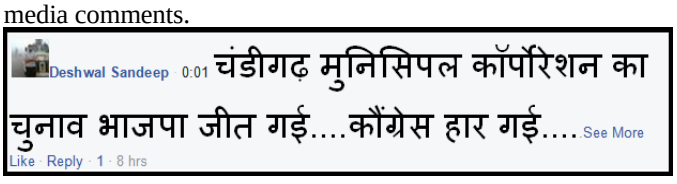

Figure 2 Text written in Single (Hindi) Language

Table 2 Distribution of Facebook Comments

	Hindi_English	*Punjabi_English*	*Average*
English	12.11%	7.87%	9.99%
Unicode(Hindi/Punjabi)	13.40%	22.47%	17.94%
Phonetic(Hindi/Punjabi)	25.52%	41.57%	33.54%
Code Mixed (Phonetic)	48.71%	28.09%	38.40%
Code Mixed (Unicode)	0.26%	0.00%	0.13%

B. *SMT Written in Bilingual/ Multilingual*

SMT are written in bilingual or multilingual using phonetic typing (roman letters) and concept of Code Mixing as shown in Fig. 3. In India people writes 38.40% comments using code mixed concept on social media as per data collected from Facebook on two topics related to Indian Politics. Only 9.99% comments are written in English Language. Table 2 shows that more than 50% people prefer Indian Languages instead of writing in English. Hindi-English code mixed comments collected from Facebook are written during the Live Speech of a Political Party and 48.71% comments are written using Code Mixed and Phonetic typing as shown in table 2. As shown in fig. 4, mostly people writes comments using Hindi-English code mixed phonetic typing in India; while in Punjab (India) preferred Punjabi language using phonetic typing.

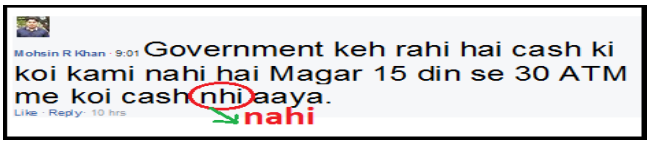

Figure 3 Text written in Bilingual

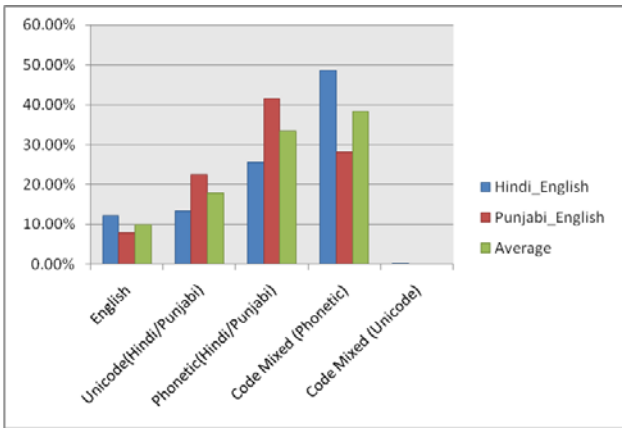

Figure 4. Comments written on Social Media (facebook)

IV. PATTERN OF CODE MIXING

Code Mixing is a term of Linguistics, which refers to using of two or more language in a conversation. There are two types of code mixing: intra-sentential and inter-sentential code mixing. When two or more language words (constituent) are mixed within a sentence is called Intra-Sentential as shown in Fig. 3, where some English words are mixed with Hindi words. Other one is Inter- Sentential code mixing (code-switching) in which one sentence is written in one language and other sentence is written in other language [14]. In this section, we focus on intra-sentential code mixing. There are three different patterns through which we explain the concept of code mixing. The three pattern of code mixing are: insertion, alternation and congruent lexicalization.

A. *Insertion*

In this case, lexical (word) term of other language is inserted into one language. In Fig. 5 (a), "a" & "b" are word (s) or constituent of "A" and "B" languages correspondingly. The word (constituent) "b" is inserted into structure of "A" language [15].

E.g. Mera **India** Mahan ha.

Here, "India" is an English word, which is inserted in between Hindi words.

B. Alternation

In this case, lexical term (word) or phrase of one language is followed by a lexical term (word) or phrase of another language. As shown in Fig.5 (b), "a" word (phrase) or constituent of "A" language is followed by "b" word (phrase) or constituent of "B" language [15].

E.g. Sarkar keh rahi hai paisa ki koi kami nahi hai, but still there is no cash in ATMs.

Here, Hindi language phrases are followed by English language phrases.

C. Congruent Lexicalization

In this case, lexical term (phrase) of any language from "A" and "B" can be inserted randomly as shown in fig.5(c) [15]. In Fig.3, the English words –government, cash, ATM are randomly inserted into Hindi words language.

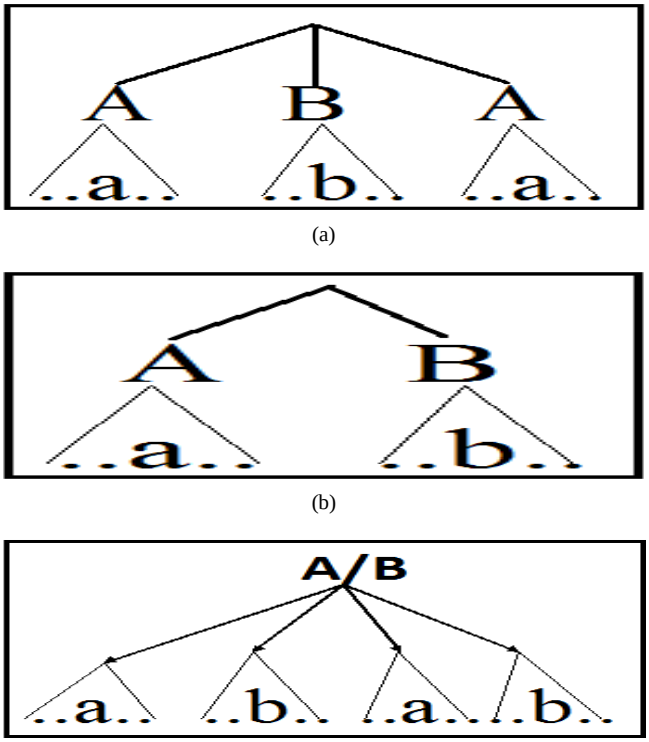

Figure 5 Different Pattern of Code Mixing: - (a) insertion, (b) alternation and (c) congruent lexicalization

V. CHALLENGES OF SOCIAL MEDIA TEXT

SMT are generally small in size like twitter messages size is 140. In Fig. 2, 3, 7, 8 the size of text comments is less than 50 words. Variation in writing style of different people causes a number of challenges which are discussed below.

A. Non Structured Data

The contents available on websites are in non structured form. There is diversity in the sources of contents from social websites, books, journals, newspapers, health records and web documents etc. The different available formats are text, video, audio, image etc. as shown in Fig.6. These diversities in the source of data and formats increase complexity for NLP.

B. Non Standard Abbreviation

A new trend is increasing on social media, that the users write comments (text) in non standard abbreviations like gm for good morning, gnt/gn8 for good night. Its variation in abbreviation of words or phrase depends on person. It can be normalized to its original words by using list of frequently used abbreviations.

C. Phonetic Similarity of Spellings

In Punjab state of India, generally people writes on social media in mix of Hindi, Punjabi and English. They write using phonetic typing. Thus there are some words, which share the same surface form [16].e.g. the word "to" is in Hindi it is "तो", in Punjabi it is "ਤੇ " and in English it is "to". To resolve this type of ambiguity is a challenging task.

Fig.6 WhatsApp message contain Video, Audio and Text data

D. Typing Errors

Anyone can frequently and correctly writes using pen and paper. But when they use keyboard to write anything then chances of typing error increases, which is a very common issue. In case of SMT, the auto correction feature does not work well, due to phonetic typing as shown in given Fig.7. The correct word is "hai" & "meediya" instead of "He" & "midia" correspondingly.

Figure 7 Shows Typing Error

E. Repetition of Characters

When writers want to emphases on a word then they simply repeats the characters within the words as shown in Fig.8. In word "picccc" the character "c" is repeated four times to make emphases on original word "pic". This is a challenging task to identify the original word from this kind of words.

Figure 8 Show Repetitions of Characters within Word

F. Cognitive Error

Cognitive errors occur when user writes wrong word in place of correct word according to that sentence in which word is used. The spelling of word is correct but only the meaning or sense of word is wrong according to the sentence [12].

E.g. I like **there** living style.

Here, word "there" is wrong word according to the sentence, the correct word is "their". This problem can be reduced or resolve with the help of bigram corpus.

G. Phonetic Misspelling

When user writes any comments on social media using phonetic typing, then there is chance of misspelling as shown in Fig. 3. But anyone can understand the meaning of sentence even there is phonetic misspelling. The problem occurs, when we use machine or software to process these texts for NLP applications. The correct spelling of "नहीं" in phonetic is "nahi" instead of "nhi" in Fig. 3.

H. Multiword Tokens

Users write a single word in place of multiple words e.g. **asap** for multiple words "**as soon as possible**". This has become new writing trend on social media to use short cut.

I. Creative Use of Punctuation

Some people on social media creatively use Punctuation to express their feeling. E.g. for happiness **:-)** and for sad **:- (** [16].

J. Non-Linguistic Sounds

The internet users express their feelings (happiness or sadness) using lexical term, which are not the part of any language. E.g. express happiness by writing – hahahaha

K. Lack of Resources:-

Most of the online available parser and POS taggers are only for English Language. There are very few NLP tools are available for Social Media Code Mixed Text and not much datasets are available in other languages (except English).

L. Multilinguality in a Sentence:-

Table 2 shows that 38.40% comments written on social media are in Bilingual. Also a Fig.3 shows that comment is written in Bilingual. It is easy to deal with content in a single language. But sometime, few words written in the content belongs to other languages than the base language. These words are considered as stop words. But many times, these words are proven valuable or meaningful words.

VI. CONCLUSION

Analysis and pre-processing of social media text is a fastest growing research field in area of NLP. Most of internet user preferred writing text/comments in their native language (using phonetic typing) instead of writing in English on social media as shown in table 2. Sentiment analysis of SMT is a new concept in field of Linguistic Computation. There are a number of challenges of SMT for pre-processing due to unstructured form and small size along with linguistic noises. Some of the challenges can be reduced or resolve, up to some extent with the help of bigram or trigram corpus, abbreviations list, word dictionary etc. Our research is in progress to resolve these challenges for Indian Language Social Media text.

VII. REFERENCES

[1] http://www.business-standard.com/article/technology/india-to-have-the-highest-internet-traffic-growth-rate-113071000014_1.html

[2] http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[3] http://www.cio.com/article/2915592/social-media/7-staggering-social-media-use-by-the-minute-stats.html#slide8

[4] K. Toutanova, D. Kleina, C. Manning and Y. Singer, "Feature-rich part-of speech tagging with a cyclic dependency network" In the Proceeding of 10th Recent Advances of Natural Language Processing (RANLP), 2015.

[5] S.K. Singh and S. Paul, "Sentiment Analysis of Social Issues and Sentiment Score Calculation of Negative Prefixes", International Journal of Applied Engineering Research, Vol. 10 No.55, pp. 1694-1699, 2015.

[6] Y. Vyas, S. Gella, J. Sharma, K. Bali and M. Choudhury, "POS Tagging of English-Hindi Code-Mixed Social Media Content". Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974−979,October 25-29, 2014, Doha, Qatar.

[7] R. Sequiera, M. Choudhury and K. Bali, "POS Tagging of Hindi-English Code Mixed Text from Social Media: Some Machine Learning Experiments". Published In ICON 2015.

[8] Raj Nath Patel, Prakash B. Pimpale and M. Sasikumar, "Recurrent Neural Network based Part-of-Speech Tagger for Code-Mixed Social Media Text", Published in ICON 2016, arXiv:1611.04989v2 [cs.CL]

[9] Souvick Ghosh, Satanu Ghosh and Dipankar Das, "Part-of-speech Tagging of Code-Mixed Social Media Text", Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 90−97, Austin, TX, November 1, 2016.

[10] Anik Dey and Pascale Fung, "A Hindi English code-switching corpus", In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 2410– 2413, Reykjavik, Iceland. European Language Resources Association (ELRA), 2014.

[11] U. Barman, A. Das, J. Wagner and J. Foster, "Code Mixing: A Challenge for Language Identification in the Language of Social Media", Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 13–23, October 25, 2014, Doha, Qatar. 2014 Association for Computational Linguistics.

[12] S.Dutta,T. Saha, S. Banerjee and S.K. Naskar, "Text Normalization in Code-Mixed Social Media Text", 2nd International Conference on Recent Trends in Information Systems, 2015, pp.378-382.

[13] Karandikar Vallabh Shankar, "A study of code mixing in selected novels in Indian English", Savitribai Phule Pune University, March, 2014, Thesis.

[14] Mary W.J. Tay, "Code-Switching and Code-Mixing as a Communicative Strategy in Multilingual Discourse", in World Englishes, Great Britain: Pergamon Press, 1989.

[15] P. Muysken, "Bilingual speech a typology of code mixing", Cambridge University Press, 2000

[16] Jagroop Kaur and Jaswinder Singh , "Toward Normalizing Romanized Gurumukhi Text from Social Media" . Indian Journal of Science and Technology, Vol 8(27), October 2015,pp.1-6.