International Journal of Engineering & Technology, 7 (2.21) (2018) 319-322

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A comparative review of the challenges encountered in sentiment analysis of Indian regional language tweets vs English language tweets

Saini Jacob Soman[1*], P. Swaminathan[2], R. Anandan[3], K. Kalaivani[4]

[1]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
[2]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
[3]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
[4]*Department of Computer Science & Engineering, Vels Institute of Science, Technology & Advanced Studies(VISTAS), Chennai, India.*
**Corresponding author E-mail:sainijacobs@gmail.com*

Abstract

With the developed use of online medium these days for sharing views, sentiments and opinions about products, services, organization and people, micro blogging and social networking sites are acquiring a huge popularity. One of the biggest social media sites namely Twitter is used by several people to share their life events, views and opinion about different areas and concepts. Sentiment analysis is the computational research of reviews, opinions, attitudes, views and peoples' emotions about different products, services, firms and topics through categorizing them as negative and positive emotions. Sentiment analysis of tweets is a challenging task. This paper makes a critical review on the comparison of the challenges associated with sentiment analysis of Tweets in English Language versus Indian Regional Languages. Five Indian languages namely Tamil, Malayalam, Telugu, Hindi and Bengali have been considered in this research and several challenges associated with the analysis of Twitter sentiments in those languages have been identified and conceptualized in the form of a framework in this research through systematic review.

Keywords: *Sentiment analysis, Indian regional language tweets, challenges in sentiment analysis, twitter sentiment analysis of English tweets.*

1. Introduction

Twitter is a rich information source for decision making using sentiment analysis. Sentiment analysis offer better decision making provided to specific individual, service or product. Sentiment analysis is referred as a method of knowing users opinions, emotions and attitude sujected towards an item which can imply to events, topics or individuals of recent trends. Sentiment analysis can be categorized under 3 levels namely sentence level, aspect level and document level. As the platform of twitter uses tweets to denote opinions in sentence form the senetence level analysis of sentiment used for examining sentiments [1]. Sentiment analysis utilizes the NLP (natural language processing) computational techniques and text analysis to operate the classification or extraction of sentiment from the reviews of sentiment. The examination of these opinions and sentiments has distributed across several sectors namely marketing, consumer information, application, books, websites and sentiment analysis becomes an essential field in decision making [2]. Million number of tweets posted everyday comprises sentiment and opinions of users in different languages [3]. The classification of sentiment can advantage firms by offering information for examining feedback of customer for products or organizing market study. Sentiment classifiers required to manage tweets in many languages to enclose a wide part of available tweets. Traditional classifiers are always specific on language and needs huge amount of work to be used to a varied language. Sentiment analysis over Twitter provides the firms an effective and rapid way to supervise the feelings of publics towards their brands. Sentiment analysis is a challenging task. Some of the essential challenges in sentiment analysis of regional language tweets are sarcasm detection [4], thwarted expression [5], negation handling [6], scarce resource language [7], subjectivity detection [8] and domain dependence [9]. Similarly there are certain challenges in sentiment analysis of English language tweets namely negation handling [10], short informal text [11], multilingual subjectivity detection [12] and microblogging data [13]. The main aim is to compare the challenges encountered in sentiment analysis of Indian Regional Language Tweets vs English Language Tweets.

2. Review of literature

Challenges encountered in sentiment analysis of indian regional languages

There are several challenges in classifying tweet into negative and positive sentiment class. Each challenges encountered in sentiment analysis of Indian regional language are described briefly.

Tamil tweets

In modern Tamil language handling negation is a challenging one for language experts and researchers who are involved in semantic and syntax analysis. There are many negative markers in modern Tamil that cannot be derived easily from an individual root of verb or positive form. The negative marker refers a word which reverses the sentence polarity. The various works have formulated different frameworks of theory to manage the Tamil negation complexities with certain restricted success level [14]. Unlike English where negation is achieved by direct way namely not the Tamil language has many negative forms inherently that cannot be derived easily from an individual root verb or positive forms. Though negation is not a part of grammatical classification it can be handled as derived forms of verbal or verb phrases. Negations generally exist as suffixes to Tamil verbs, which makes sentiment analysis a complex process.

Twitter's informal nature leads to huge number of sentiments being posted and this has made twitter a gold mine for sentiment analysis. Several systems have used twitter as a corpus for sentiment analysis. It is essential to know people sentiment and their concerns from users tweeted post. It can be useful in facilitating the requirements of those influenced by disaster. The tweets in the actual form involve several slang words and grammatical errors because of tweets informal nature. Morphological richness of Tamil tweets leads to degradation in performance when the length of tweet is much critical [15]. As a Dravidian language Tamil has a rich morphological structure which is agglutinative. The words of Tamil are comprised of lexical roots followed by more than one affixes mainly suffixes. This study has developed a speech tagging system parts to manage verbs and nouns. So predicting a word in a language like tweets is difficult. The major barrier [16] in examining the opinions of twitter is how to manage with informal dialect used on this platform because it comprises expressions namely abbreviations, acronyms, misspelled words and slang words that are not examined in traditional media, A domain specific tag considers the entire non English words and slang words and matches those words to dictionary in Tamil to develop the classification model accuracy. This study have examined the tweets and developed a dataset in addition to dictionary in Tamil to represent the entire non English words and slang words to predict the sentiments within the tweets.

Malayalam tweets

There is a rule based [17] approach for retrieving sentiments from movie reviews of Malayalam. Negation handling is a challenging one in sentiment analysis because negations can be denoted in different ways even without the negative word usage. They mention that extraction of sentence level is efficient in movie websites the user comments are just individual sentences. The rules of negation are applied for examining the sentiment which reduces the options of error occurrence.

A novel hybrid approach [18] based on maximum entropy-classifier for sentiment analysis of Malayalam movie reviews perform Malayalam movie review sentiment classification acquired from the user as positive, neutral and negative. A hybrid method is used that is an integration of maximum entropy method which is utilized for tagging and some rules for managing special cases. The classification of maximum entropy predicts which class the movie reviews belong to a context so that it expand the clarification system entropy. Here certain norm is used to manage special cases which involve negation, dilators, intensifiers, etc. Another major challenge is that varied users spell similar entity in varied types. Sentence compression [19] is one of the challenges in sentiment analysis. When humans generate document summaries they do not retrieve sentences and integrate them. Rather they make new sentences that are with grammar mistakes that combine with one another and that seizes the essential pieces of data in the actual document. Given that greater collection of abstract or text set feasible it is possible to envision algorithms that are trained to mimic this method. In this study the author concentrates on compression of sentence an easier version of bigger challenge. This study accomplishes two targets simultaneously and their compression must be with grammar mistakes and they must acquire the most essential piece of data.

In one of the research studies [20] the beginning step is to collect corpus from novels in Malayalam. In order to avoid mistakes in grammar sentences are typed manually. As much as 100 sentence are gathered in the beginning word used inside the sentence are manually tagged as adverb and adjectives. These adverb and adjective in sentence refer the sentence emotion. Negative or positive word may have similar meaning and sarcasm sentence without or with sentiment word are critical to manage. Sarcasm is not so similar in customer review about services and products but it becomes difficult to manage with sentences. Since the language of Malayalam is agglutinative highly with rich morphology it has vast number of fluctuated words that denotes similar meaning.

Telugu tweets

Telugu is rich morphology [21] which is very critical and is one of the challenges to evolve syntactic parsers for these kinds of languages. To build a dependency parser it needs a better morphology based POS (part of speech) tagger. In this study part of speech tagger is used which has a better performance on sentences of Telugu language. This study explains the steps to develop the parser dependency for Telugu language. With the developing quantity of text feasible online and multilingual parallel sentences from sources such as government documents, multilingual websites and huge human translation archives of news, book and un-annotated parallel text is available vastly. This parallel information can be used to separate languages and transfer data from rich resource to poor resource language.

Hindi tweets

Projection of polarity [22] and sentence level translation is much challenging for the Hindi English pair as Hindi is a translation cost and free order language that adds up to error rate. In Hindi language similar word with similar meaning can exist with varied spellings so it is difficult to have entire existence of such words in a lexicon and while training a model it is quite difficult to manage all the variants of spelling. There is lack of adequate tools, annotated corpora and tools for Hindi language. Hindi is one of the mainly spoken languages in the globe [23]. Hindi is written Devanagiri script and the Hindi data is present in electronic form. Hindi is a scarce resource language where parsers are not that effective and the data requires to be tagged manually for each task due to the unavailability of well annotated standard corpora. Every language put forward challenges to be faced in terms of its semantic and syntactic structures. Hindi is a free order language with different variants of morphology, variance in spelling, variances in context and word sense ambiguity. In Hindi sentiment analysis is explored less so there is scarcity of tools and resources.

The occurrence of misspellings, poor structure of grammar, acronyms, slang and emoticons are coon and makes the sentient analysis task from these texts much critical [24]. Sentiment analysis becomes much challenging in twitter text case when user attempts to project their sentient using 140 characters. The scarcity of sentiment lexicons available for Indian language tweets is another challenge in sentient analysis. A better number of sentiment lexicons feasible for Hindi, Tamil and Bengali languages are gathered from plain texts. In a sentence the arrangement of word plays an essential part in recognizing the text subjective nature. Hindi is a free order language where the verb, subject and object can exist in any order. The arrangement of word play an essential role in determining the text polarity where the similar words set with small changes in the word arrangement could influence the polarity aspect.

Bengali tweets

The subjectivity detection [25] is about understanding if the content comprises opinions and personal view as opposed to factual data. Subjective expressions are due to the experience or culture of a community or user and hence can be specific and localized to society. Subjectivity is learnt before sentiment analysis is performed since it is important to filter out factual information to have a better understanding of problems that are shared from netizens. It is challenging to categorize subjective tweets with philosophical thoughts. This is because certain phrases become weakly subjective and hence useable on both subjective and neutral tweets. Besides that the phrases with polarity and word sense ambiguation are predicted to be prone to error.

Negation norms may be varied for various languages and hence cause unwanted mistakes. The biggest challenge with the present state of Bengali sentiment analysis [26] are the absence of a standard and big enough set of data to compare against makes the research work comparison difficult and another challenge is that none of the Bengali sentiment analysis research considers the practical perspective of the Romanized Bengali usage. The subjectivity analysis [27] cannot be achieved without an annotated corpus or lexicon. Even though several scarce resource languages have restricted resources feasible an initial lexicon or annotated dictionary is still required before a classifier with enough accuracy can be accomplished. Although the words of negation do not denote any sentiment they influence the overall tweet sentiment. The negation use in Bengali language is varied from that of English language. Unlike English language where the negative words generally exists in the mid of a sentence, the sentences of Bengali comprises negation towards the end.

The below Table I show the challenges encountered in sentiment analysis of Indian regional language Tweets:

Table I: Challenges Encountered in Sentiment Analysis of Indian Regional Language Tweets

Challenges	Description	Language
Sarcasm Detection	Sarcasm is a special kind of sentiment which defines the inverse meaning of what people express in the text. It is always denoted using intensified positive or positive words. Posting Sarcastic messages on social media becomes a new fashion to avoid negative words. Sarcasm is a special type of sentiment that generally flips the orientation of the view in a given text piece. The sarcasm sentence generally looks positive but the overall meaning indicates negative due to the existence of sarcasm.	Hindi
Thwarted Expression	There are certain sentences wherein small number of text decides the documents overall polarity. Easy bag of word approach will fail in these cases as several words used are positive but the main sentiment is negative because of the critical last sentence. In traditional classification of text this would have been categorized as positive as the phrase frequency is much essential than the phrase presence.	Nil
Negation Handling	Handling negation is one of the largest challenges in sentiment analysis. This technique generates a solution to develop negative evaluation with the developed BOW (bag of words) technique. This challenge is classified into implicit and explicit negation. Implicit negation is the unconscious level which are formed involuntarily and are unknown typically to users without any negative keywords. Explicit negation is formed deliberately and are simple to self report by keywords.	Malayalam, Tamil
Scarce resource language	In sentiment analysis the scarce resource language is one of the major challenges. The languages for which the annotated corpus, availability of tools and resources is restricted and under the phase of development. This infers to those with a general dictionary available and/or lacking of developed resources of text processing. The different language variants used on social media belong to this category. The bottleneck for carrying out sentiment analysis is because of benchmark datasets non availability and scarcity of tools and resources.	Telugu, Hindi
Subjectivity detection	Subjectivity detection is used to make a difference between non opinionated and opinionated sentences. A subjectivity detection method can be involved in the system to predict the objective facts. Thus the system performance can be raised. But this is always critical to perform. Subjectivity detection is a very difficult challenge for machines with restricted emotional capabilities and also for human beings. Subjectivity detection is concerned with whether the sentiment expressed is associated to the similar concept or satisfies the overall target of sentiment analysis system.	Bengali
Domain Dependence	The sentiment analysis biggest challenge is the sentiment words domain dependent nature. There are several words whose polarity alters from one domain to another domain. The same phrase or sentence can have varied meanings in varied domains. One set of features may provide better performance in one domain at the same time it operate bad in some other domain.	Nil

Challenges encountered in sentiment analysis of English Tweets

Negation permits to change a word's meaning to its inverse meaning [28]. Therefore during the extraction of feature it is essential to represent the process whether or not a word is negated. If the negation is not handled the algorithm will understanding the inverse meaning of the phrase and will have reduced accurate predictions. The research multilingual data can develop by 5 percent the subjectivity classification performance in English language [29]. The author found that a perfect sense to sense representation between languages is not possible as a specific sense may indicate extra uses and meanings in one language compared to another language. However they also offer proof that a multilingual feature space is capable to depend on double co-existence metrics studied from definitions of equivalent sense thus permitting for a much robust modeling than when regarding every language separately.

In micro blogging surroundings [30] the real time communication is a major feature and the ability to examine data automatically examine sentiments of user as discussions evolve is a challenging problem. The challenges which analysis of data has to manage in micro blogging data case is the use of abbreviated, informal and developed language as well as lack of data due to short messages that are transformed. Nowadays micro blogging has become a familiar tool of communication among online users [31]. Micro blogging sites are rich data sources for sentiment analysis because in examining short informal texts namely comments, tweets or blogs it turns out that emoticons offers a difficult piece of data. However emojis have not been used and no resource with emoji sentiment data has been offered. Emoticons have proved difficult in automated sentiment informal texts classification. The below Table II shows the challenges encountered in sentiment analysis of English language Tweets:

Table II: Challenges Encountered in Sentiment Analysis of English Language Tweets

Challenges	Description
Negation Handling	Negation was managed using a predefined type of negation tokens then the prefix was attached to the following token until a punctuation mark of clause level is annotated in a negated context. A list of forty five negative words and phrases was used to signal the negation.
Short Informal Text	Short informal text is one of the challenge in sentiment analysis. They are restricted in length generally spanning one or less than one sentence. They tend to have several slang phrases, misspellings and shortened word forms. They also have special markers namely hashtags that are used to facilitate search but also represent a sentiment or topic.
Multilingual subjectivity detection	Multilingual data can develop by the subjectivity classification performance in English. A perfect sense to sense representation between languages is not possible may indicate extra uses and meanings in one language compared to another language. A multilingual feature space is capable to depend on double co-existence metrics learned from definitions of equivalent sense thus permitting for a much robust modeling than when regarding every language separately. Boosting one view/language develops the performance for classification of subjectivity with respect to multilingual types.
Microblogging data	Microblogging sites have developed to become a source of different type of data. This is because of microblogs on which people post real time messages about their views on different concepts, discuss present problems, express and complain positive sentiment for products which they utilize in daily life. Companies promoting their products have initiated to poll these microblogs to acquire a sense of usual sentiment for their product. Several times these firms learn reactions of user and response to users on microblogs. This challenge can be resolved by building technology to summarize and detect an overall sentiment.

3. Findings and research framework

It is clear from the above literature that the challenge in twitter sentiment analysis is in Indian regional languages tweets is multifold as well as greater when compared with that of the English language tweet analysis. Fig 1 depicts challenges in twitter sentiment analysis by comparing Indian regional languages with English language.

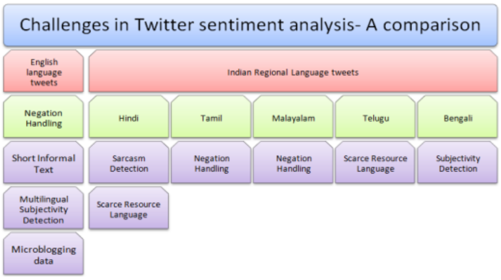

Fig. 1: Challenges in twitter sentiment analysis-A framework of comparison between english and indian regional language Tweets (Source: Author)

From the above figure the challenges in English language twitter sentiment analysis are negation handling, short informal text, multilingual subjectivity detection and microblogging data. Similarly the challenges of twitter sentient analysis for the Indian regional languages namely Hindi, Tamil, Malayalam, Telugu and Bengali are sarcasm detection, thwarting expression, negation handling, scarce resource language, subjectivity detection and domain dependence.

4. Discussion and conclusion

Sentiment analysis plays an essential role in recognizing the views, emotional states and attitudes of users and sentiment analysis in Twitter is an active research area. Though several sentiment analysis researches have been discussed on English languages using main platforms namely official or news documents the increasing attention is placed on social media content analysis to facilitate the understanding of the community well being or the perceived image of a product or firm. Major challenges in both the Indian regional and English languages have been identified and compared in this study. It can be concluded that Tweets in English language could be analyzed for its sentiments with lesser difficulty when compared with that of Indian Regional Languages. The study could be further extended by developing an algorithm for analyzing the Tweets in both Indian regional as well as English languages and testing them on real-time tweets in the future.

References

[1] Venugopalan M & Gupta D, "Exploring sentiment analysis on twitter data", *Proceedings of eighth International Conference on Contemporary Computing*, (2015), pp.241-243.

[2] Chalothom T & Ellman J, "Simple approaches of sentiment analysis via ensemble learning", *Information science and applications*, (2015), pp.631-639.

[3] Narr S, Hulfenhaus M & Albayrak S, "Language-independent twitter sentiment analysis", *Knowledge discovery and machine learning (KDML)*, (2012), pp.12-14.

[4] Riloff E, Qadir A, Surve P, De Silva L, Gilbert N & Huang R, "Sarcasm as contrast between a positive sentiment and negative situation", *Proceedings of Conference on Empirical Methods in Natural Language Processing*, (2013), pp.704-714.

[5] Kaur J, "A Review Paper on Twitter Sentiment Analysis Techniques", *International Journal for Research in Applied Science & Engineering Technology*, Vol.4, No.10, (2016), pp.61-69.

[6] Remus R, "Modeling and Representing Negation in Data-driven Machine Learning-based Sentiment Analysis", *ESSEM@ AI* IA* , (2013), pp.22-33.

[7] Irvine A & Callison-Burch C, "Combining bilingual and comparable corpora for low resource machine translation", *Proceedings of the eighth workshop on statistical machine translation*, (2013), pp. 262–270.

[8] Severyn A & Moschitti A, "Twitter sentiment analysis with deep convolutional neural networks", *SIGIR*, (2015), pp. 959–962.

[9] Xiang B & Zhou L, "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol.2, (2014), pp.434-439.

[10] Ghaleb OAM & Vijendran AS, "The Challenges of Sentiment Analysis on Social Web Communities", *International Conference on Intelligent Computing and Technology*, (2017), pp.21-29.

[11] Agrawal A & An A, "Kea: Sentiment Analysis of Phrases within short texts", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, (2014), pp.380–384.

[12] Makrynioti N & Vassalos V, "Sentiment extraction from tweets: multilingual challenges", *International Conference on Big Data Analytics and Knowledge Discovery*, (2015), pp.136-148.

[13] Bahrainian SA & Dengel A, "Sentiment analysis and summarization of twitter data", *IEEE 16th International Conference on Computational Science and Engineering*, (2013), pp.227-234.

[14] Asmi A & Ishaya T, "Negation Identification and Calculation in Sentiment Analysis", *Proceedings of the Second International Conference on Advances in Information Mining and Management*, (2012), pp.1-7.

[15] Sharmista A & Ramaswami M, "Tree Based Opinion Mining in Tamil for Product Recommendations using R", *International Journal of Computational Intelligence and Informatics*, Vol.6, No.2, (2015), pp.110-119.

[16] Bravo-Marquez F, Frank E & Pfahringer B, "From unlabelled tweets to twitter-specific opinion words", *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2015), pp.743-746.

[17] Nair DS, Jayan JP & Sherly E, "Sentiment Analysis of Malayalam film review using machine learning techniques", *IEEE International Conference on Advances in Computing, Communications and Informatics*, (2015).

[18] Anagha M, Kumar RR, Sreetha K & Reghu Raj PC, "A Novel Hybrid Approach on Maximum Entropy Classifier for Sentiment Analysis of Malayalam Movie Reviews", *International Journal Of Scientific Research*, Vol.4, (2015).

[19] Anu PC, Athila M, Heera BM, Lini KU & Cheerotha LR, "Aspect Based Sentiment Analysis in Malayalam", *International Journal of Advances in Engineering and Scientific Research*, Vol.3, No.6, (2016), pp.27-34.

[20] Shankar R, Shilpa KM, Patil S & Swamy S, "A Survey on Sentimental Analysis in Different Indian Dialects", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol.5, No.4, (2016), pp.1072-1076.

[21] Nagaraju G, Mangathayaru N & Rani BP, "Dependency Parser for Telugu Language", *Proceedings of the Second ACM International Conference on Information and Communication Technology for Competitive Strategies*, (2016), pp.138-139.

[22] Rai V, Vijay S & Sharma DP, "A Karaka Based Approach to Cross Lingual Sentiment Analysis", *International Journal of Languages, Literature and Linguistics*, Vol.3, No.4, (2017), pp.226-229.

[23] Mishra D, Venugopalan M & Gupta D, "Context Specific Lexicon for Hindi Reviews", *Procedia Computer Science*, (2016), pp.554-563.

[24] Patra BG, Das D, Das A & Prasath R, "Shared task on sentiment analysis in Indian Languages (sail) tweets-an overview", *International Conference on Mining Intelligence and Knowledge Exploration*, (2015), pp.650-655.

[25] Cambria E, Olsher D & Rajagopal D, "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis", *Proceedings of AAAI conference on Artificial Intelligence*, (2014), pp.1515–1521.

[26] Khan S, "Convergence in spelling, and spell-checker for Romanized Bangla in computers and mobile phones", *IEEE International Conference on Informatics, Electronics & Vision (ICIEV)*, (2014), pp.1-5.

[27] Chowdhury S & Chowdhury W, "Performing sentiment analysis in Bangla microblog posts", *the IEEE International Conference on Informatics, Electronics & Vision (ICIEV)*, (2014), pp.1-6.

[28] Narayanan V, Arora I & Bhatia A, "Fast and accurate sentiment classification using an enhanced Naive Bayes model", the *International Conference on Intelligent Data Engineering and Automated Learning* Springer, (2013), pp.194-201.

[29] Banea C, Mihalcea R & Wiebe J, "Sense-level subjectivity in a multilingual setting", *Computer Speech & Language*, Vol.28, No.1, (2014), pp.7-19.

[30] Karanasou M, Ampla A, Doulkeridis C & Halkidi M, "Scalable and Real-time Sentiment Analysis of Twitter Data", *16th IEEE International Conference on Data Mining Workshops (ICDMW)*, (2016), pp.944-951.

[31] Bilgin M & Şentürk İF, "Sentiment analysis on Twitter data with semi-supervised Doc2Vec", *IEEE International Conference on Computer Science and Engineering*, (2017), pp.661-666.