Sentiment Analysis of Products Reviews Containing English and Hindi texts

Jyoti Prakash Singh[1], Nripendra P. Rana[2] and Yogesh K. Dwivedi[2]

[1] National Institute of Technology Patna, Bihar, India
[2] School of Management, Swansea University, UK, SA2 8PP

Abstract. The online shopping is increasing rapidly because of its convenience to buy from home and comparing products from their reviews written by other purchasers. When people buy a product, they express their emotions about that product in the form of review. In Indian context, It is found that the reviews contain Hindi text along with English. It is also found that most of the Hindi text contains opinionated words like *bahut achha*, *bakbas*, *pesa wasool* etc. We have tried to find out different Hindi texts appearing in product reviews written on Indian E-commerce portals. We have also developed a system which takes all those reviews containing Hindi as well as English texts and find out the sentiment expressed in that review for each attribute of the product as well as a final review of the product.

Key words: Sentiment Analysis; POS-Tagging; Review Analysis; Product Summarization.

1 Introduction

The life style of society is changing with the penetration of Internet and E-commerce in every corner of the world. The advertisement and friends recommendations were a major source of information while buying a product. The number of recommendations was a limited one to compare similar products of different brands. Nowadays, as the e-commerce business has grown up they are offering more products. The e-commerce websites also requests their customers to write their experience about the product they brought in the form of a product review. These reviews offer significant information to buyers about the product they are planning to buy and also enable them to compare products of different brands. The reviews help consumers to choose the best products by comparing them based on other consumers evaluation of the products. It also aids in the improvement of the product by informing the manufacturers about the advantages and defects of their products. The number of reviews about the products grows with the growth of e-commerce businesses. It becomes very difficult for buyers and sellers to manually analyze a large number of reviews and get any meaningful information. This attracts a lot of researchers to automate the analysis of reviews and get valuable information hidden in the reviews [3, 5].

The reviews written by Indian buyers are mainly in English, but it contains some Hindi texts (written in English Scripts only) also as Hindi is a prevalent language in India. Some of the most widely used Hindi words are *bahut achha*, *bakbas*, *pesa wasool* as found in a number of reviews. Most of these words are opinionated and contain strong opinions either good or bad. Most of the earlier work done in the area of finding polarity of opinions for product reviews neglect these texts as they are mainly developed for English texts only. According to the best of our knowledge, no work has been reported which consider the correction of these typos and includes the sentiment of the Hindi words together with English texts.

In this work, we have proposed a sentiment analysis system which works for reviews contain English as well as Hindi opinionated text. First of all we have gathered possible Hindi opinionated text from reviews appearing on Indian popular E-commerce sites such as *amazon.in, flipkart.com, snapdeal.com, shopclues.com* and so on. These Hindi texts are preprocessed and their equivalent English words are found. The summarized review of the product is then calculated consulting sentiwordnet database.

The rest of the paper is organized as follows: The proposed system architecture and algorithm are discussed in section 2. In section 3, we present our results and finally in section 4, we conclude the paper.

2 Proposed Work

Product reviews from popular Indian e-commerce sites like *amazon.in, flipkart.com, snapdeal.com, shopclues.com* contains are collected as our dataset. The dataset has a lot of typos in the form of joint words like *verygood* as well as abbreviations containing numerals such as *gr8 for great*. The dataset also contains Hindi words like *"bakbas", "bekar", "achchha"* (written in English script only). Some sample typos gathered from various Indian e-commerce websites are given in Table 2 and 3. Table 1 contains words of Hindi Texts typed in English along with their English equivalent text. Table 2 contains some popular abbreviations used online for review, chatting, etc. along with their correct form in English. Some joint words (missing space) are shown in Table 3.

Table 1. List of Wrong Words

Wrong Words	Corrected Words
ye	this
achha	good
g8t	great
n8t	night
h	is
som1	some one

Table 2. List of Wrong Words

Wrong Words	Corrected Words
Gud	good
gooood	good
exclent	excellent
bd	bad
awesm	awesome

Table 3. List of Joint Words

Jointed Words	Corrected Words
verygood	very good
verybad	very bad
bahutbura	bahut bura

One of the primary focuses of in this work is to pre-process the product review available on Indian E-commerce sites so that the reviews contain only English Text. Once reviews are converted to English text, Part of Speech (POS) Tagging to the text is done using wordnet [4] database. Once POS tagging is done, the adjective, noun, and adverb are extracted. Further, sentiwordnet [2, 1] database is used to assign numerical values to the adjectives contained in the review. The proposed system architecture is shown in Figure 1.

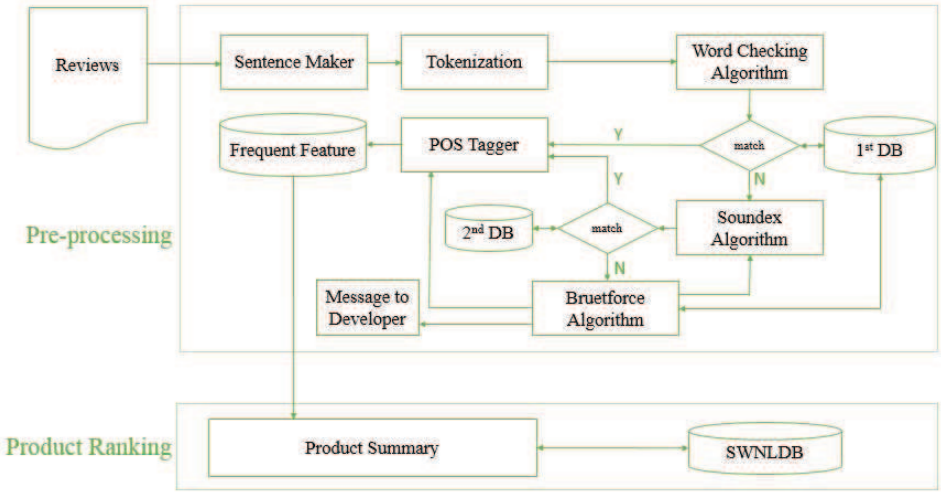

Fig. 1. Proposed System Architecture

A pseudo code for our proposed system is given below.

Proposed Algorithm:
Step 1: tokenize based on space.
Step 2: Consult wordnet

1. If word is matched, then goto POS tagger.
2. else correct word and go to POS Tagger

Step 4: Noun, Adverb, and Adjective are stored in frequent feature database.
Step 5: Generate the product summary with the help of SentiWordNet Lexical databases.

The working of our scheme is traced with the aid of an example presented here. *Yeh achha camera h. eski pictre quality bahut achhi hai. The pics resolution is enough. Zoom is bakbas. focus is verygd bt not g8t.*

The document (Complete review) is broken down into several sentences based on [.], [?], And [!] Mark. For example review, sentences are:
S1. Yeh achha camera h.
S2. eski pictre quality bahut achhi hai.
S3. The pics resolution is enough.
S4. Zoom is bakbas.
S5. focus is verygd bt not g8t.

Yeh is a Hindi word whose English equivalent is *This*. *Achha* is another Hindi word meaning *good* in English. The complete review is written in English after correcting and converting every word to English.

s1. This is a good camera.
s2. Its picture quality is very good.
s3. The picture resolution is enough.
s4. Zoom is bad.
s5. Focus is very good but not great.

The POS tagging is applied as given below for just one sentence. We have used Penn Treebank tagset for Part of Speech Tagging

This========= [This_DT]
Good========= [Good_JJ]
Camera======= [Camera_NN]
Is========== [Is_VBZ]
. ==========[_UH]

Next step is to consult the sentiwordnet database to find the priority of the adjective to find the sentiment value of the sentence. The score of every adverb and adjective are given in the SentiWordNet lexical database. We have listed here some of them given in Table 4 which are going to be used in the above example. Where nil represents the neither positive nor negative orientation of words. And negation represents the multiplier factor which having -1 value.

For above example, sentence s1 has good as adjective whose sentiment score is +0.75. In this sentence there is no adverb or negation, so the sentiment score for sentence s1 is +0.75. For second sentence s2, there is very (adverb) with the

Table 4. Score List of Words

Word	Orientation of Word	Score
Good	positive	.75
Great	positive	.875
Awesome	positive	.875
Excellent	positive	1
Well	positive	.75
Average	positive	.375
Enough	neutral	.875
Bad	negative	.65
Very	nil	.5
Not	negation	-1

adjective, so the sentiment score of s2 is 1.25 which is a sum of scores of good (0.75) and very (0.5). The sentiment score of each sentence is shown in Table **??**

Table 5. Sentence Wise Score

Sentence Number	Score	Score Type
S1	0.750	Positive
S2	1.250	Positive
S3	0.125	Positive
S4	0.625	Negative
S5	0.275	Positive

The polarity of the review of a product is determined by finding the polarity of each feature of the product across all reviews and finding a weighted sum of all features.

3 Result

We have collected 1100 reviews from *flipkart.com* and *amazon.in* of three popular mobile brands in India at the time of writing this paper. The results show that for Android based smart-phone people are talking about features like *camera, battery, memory, processor, RAM, display, price, weight and phone*. Out of these features battery, camera and display are found to be more prominent one across all phones. The results are shown in Table 6,7 and **??**. A zero (0) in both positive and negative row shows that no one has given any opinion about that feature of that product.

4 Conclusion

We have designed a sentiment analysis system which can take reviews written in Hindi as well as English texts and find the sentiment of customers for that

Table 6. Summary of the Review of product **Samsung Galaxy S3 Neo**

Phone Feature	Scores Type	Percentage(%)
Camera	Positive	100
	Negative	0.0
Battery	Positive	61.76
	Negative	38.23
Memory	Positive	100
	Negative	0.0
Processor	Positive	100
	Negative	0.0
RAM	Positive	100
	Negative	0.0
Display	Positive	90.5
	Negative	9.5
Price	Positive	0
	Negative	100
Weight	Positive	100
	Negative	0.0
Phone	Positive	55.07
	Negative	44.93
Overall	**Positive**	**68.86**
	Negative	**31.134**

product. We have taken a dictionary based approach to correct the wrong words and replace Hindi text with their English equivalent. We further want to extend this system with a machine learning algorithm to correct the wrong words and Hindi words. The final opinion score is a weighted average of all the features of the product under consideration. We also are working to identify most prominent features of a product to calculate the final opinion score.

References

1. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.
2. Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422, 2006.
3. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.
4. George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.
5. Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology*

Table 7. Summary of the Review of product **Asus Zenfone 2**

Phone Feature	Scores Type	Percentage(%)
Camera	Positive	66.39
	Negative	33.61
Battery	Positive	90.82
	Negative	9.18
Memory	Positive	100
	Negative	0
Processor	Positive	100
	Negative	0
RAM	Positive	100
	Negative	0
Display	Positive	0
	Negative	100
Price	Positive	0
	Negative	0
Weight	Positive	0
	Negative	100
Phone	Positive	89.09
	Negative	10.91
Overall	**Positive**	**75.14**
	Negative	**21.15**

and Empirical Methods in Natural Language Processing, HLT '05, pages 339–346, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Table 8. Summary of the Review of product **Honor 4X**

Phone Feature	Scores Type	Percentage(%)
Camera	Positive	60.17
	Negative	39.82
Battery	Positive	66.09
	Negative	33.9
Memory	Positive	49.24
	Negative	50.76
Processor	Positive	49.25
	Negative	50.74
RAM	Positive	48.24
	Negative	51.75
Display	Positive	100
	Negative	0.0
Price	Positive	0
	Negative	0
Weight	Positive	0
	Negative	0
Phone	Positive	62.02
	Negative	37.98
Overall	**Positive**	**62.60**
	Negative	**32.26**