Improvisation of Experience of Indian Railways using Sentimental Analysis

Anmol Rai Gupta
B.Tech. SCSE
VIT University, Vellore (India)

L. Shalini
Assistant Professor (Sr.), SCSE
VIT University, Vellore (India)

ABSTRACT

Over the past decade Indian Railways is the biggest public sector enterprise in India. Market is becoming more users oriented these days and customers' feedbacks have taken a central place. Thus, in this paper sentimental analysis has been performed on the commuters' feedback by using the concepts of polarity dictionary and sentimental orientation calculation. Each sentence has been allotted with a score calculated by mappings from the polarity dictionary, with the scores varying within a fixed pre-defined range. Thus the score gives an insight into the feedbacks of the customers and hence, the experience of commuters and railway industry revenue generation can be improved by using this study.

Keywords:

Indian Railways, Sentimental analysis, Polarity dictionary, Sentimental Orientation Calculation, Natural language processing.

1. INTRODUCTION

Indian Railways have emerged as the world's largest rail network. In the present India, almost 20 million travel through Indian Railways on daily basis. With that number, it is no doubt the biggest public sector enterprise in India; leading to involvement of billions of rupees. As with the globalization, market has become more users oriented, Indian Railways too aims at its customer's satisfaction. Indian Railways started taking written feedback from the travelers in almost all the major long distance trains. Since the web has taken central place in most of the people's lives these days, there are many blogs, forums, discussion groups etc. where travelers post their experiences and discuss about the various issues linked with the trains. With the outset of such blogs, sentimental analysis approach is growing in every industry for various product reviews and service reviews.

In this paper, the research has been based on the data procured from blogs and forums. Since the posts on these sites are given by travelers, thereby fall under the category of user generated content. In the research carried out in this paper, polarity dictionary generation is proposed along with the allotted intensity to the various seed words and quantifiers so as to analyze the experience of various commuters of Indian Railways.

The paper is constructed below. In Section 2, literature background has been given. Section 3 explains the proposed method of sentimental analysis. Section 4 consists of observations along with their results and discussions. Finally in Section 5, conclusions and future works have been described.

2. LITERATURE BACKGROUND

Various researches have been done using sentimental analyses. These researchers can be classified into broadly two categories: one based on machines (eg Support Vector Machine) and others based on the polarity dictionary and corpus development. Fuji et al. [1] conducted a research on sentimental influence on twitter. With the development on polarity dictionary, they formulated score for different tweets. Also, they calculated the influencing probabilities of users using Bernoulli distribution and thus concluded that there's high correlation between Twitter users' influencing probabilities and influenced probabilities.

Keisuke et al. [2] analyzed stock market with semi supervised learning. In their study they determined sentimental polarities of stock market news using a polarity dictionary, thereby making it easier for investors in making the appropriate decision. In their study they have used bootstrapping method to construct a small polarity dictionary containing some seed words. They compared the frequency of positive words with negative words thereby deciding the polarity of the article on the whole.

Feifan Liu et al. [3] examined various linguistic features for sentimental and used them to do comparative study on blogs and review data. For determining the polarity of whole document the authors used supervised learning framework and analyzed various features. The authors observed that performance of blogs was actually worse than the reviews and thus suggested two methods to improve the blog polarity classification: a) extract only topic relevant data b) apply adaptive methods using review methods. They noticed a significant improvement in the performance of the blogs by using suggested methods.

3. PROPOSED METHOD

In this paper the concept of polarity dictionary for the sentimental analysis has been used. For that the polarity dictionary first using the seed words and then that dictionary has been expanded using the words that occur most frequently in the feedbacks. Also, since the opinions are highly subjective, therefore just classification of polarities for different words won't give the accurate results. Thereby the concept of intensity of words was implemented because every emotion has intensity. Different points were given to various polar words: ranging from -3 to +3; -3 indicating highly negative and +3 indicating highly positive word. Besides that percentage shares have been given to various quantifiers so that if someone mentions that "the food was very good", then the score for this sentence should be different from that of "the food was good". In the proposed method the concept of root mean square was implemented for calculating the score as the points can be negative as well as positive and root mean square is very much useful in such cases when variates are positive and negative.

For a particular sentence we calculate the score as follows.

$$S(i) = \frac{P(1)+P(2)+P(3)+\cdots+P(N)}{\sqrt{\frac{1}{N}}(P(1)^2+P(2)^2+P(3)^2+\cdots+P(N)^2)}$$

Where P (i) indicates the points of i^{th} polar word along with the quantifier attached to it (if any).

And S (i) indicates the score for the i^{th} sentence.

For a post which has multiple sentences we compute the final score as follows:

$$S= \frac{S(1)+S(2)+S(3)+\cdots+S(N)}{\sqrt{\frac{1}{N}}(S(1)^2+S(2)^2+S(3)^2+\cdots+S(N)^2)}$$

So, if the score is positive it means a positive feedback, if it's negative then it indicates that the feedback is negative else the feedback is neutral. Also, the degree of positivity or negativity can be decided by looking at the deviation from 0.

3.1 Opinion Retrieval

Whenever the opinions are to be procured from user generated content then it leads to various problems and the retriever should be cautious while doing so. In user generated contents, there's no restriction on the contents being posted and also the parameters can be really vast. There are many opinions which can be categorized under Opinion Spam (Bing Lui.2010 [4]). For the purpose of this research, all those posts have been categorized under opinion spam which are either non opinions e.g. various pictures of trains, videos of running trains, waiting list status' posts etc. or in which the sentiments expressed are ambiguous.

Thereby, the fetched posts broadly fall into two main categories: Direct Opinions and Comparative Opinions: Direct opinions are those in which a person has expressed sentiments about various features of train journey: food, quality of linen, hygiene etc. along with some adjectives and quantifiers. Comparative Opinions are those in which a person has compared one feature of the train with some other train e.g. "The food of this train is not as good as Delhi Bangalore Rajdhani"; thereby expressing a negative emotion. Only those comparative opinions have been considered in which the comparison is made with the selected 4 trains only for which we have based our analysis; thus ensuring the domain.

3.2 Opinion Lexicon Generation: Quantifiers and Polar words

This section specifies the quantifiers that have been considered along with their percentage share in the polarity scores of the polar words. Also, classification has been done for some polar words on the basis of their polarity scores. When the quantifiers are used along with the polar words, then they add or subtract the given share to the polarity, thereby expressing the intensity. The percentage share of various quantifiers and the points of various polar words have been given in Table 1 and Table 2 respectively.

Table1. Quantifiers with percentage share

Quantifier	Share
Very	+30%
Extremely	+60%
Not	-20%
Extra	+40%
Throughout	+20%
Quite	+25%
Much	+20%
Heavily	+60%
More	+40%

Table 2. Sample words in polarity dictionary with respective polarity

-3	-2	-1	0	+1	+2	+3
Disastrous	Terrible	Bad	Okay	Good	Superb	Best
Pathetic	Dismal	Unhygienic	Fine	Clean	Awesome	Marvelous
Horrible	Rude	Dirty	Suitable	Tasty	Punctual	Excellent
Worst	Damaged	Poor	Appropriate	Lovely	Wow	Fast
	Complaint	Disappointing		Nice	Great	Regal
		Late				
		Sad				

4. OBSERVATIONS, RESULTS AND DISCUSSIONS

The proposed strategy was implemented on 4 Rajdhani trains. : Hazrat Nizamuddin- Chennai Central Rajdhani Express (12434), Howrah New Delhi Rajdhani Express (12302), Hazrat Nizamuddin- Bangalore Rajdhani Express (12429) and Mumbai Rajdhani Express (12952). Some of the sample scores that have been calculated are as follows:

1) Super pathetic train...horrible food..coach shakes wildly when doing 110 kmph+......pantry people rude...

S= -2.90

2) "Food: This time Ok. Snacks and lunch/dinner are just Okay . Pantry : Not good initially. They don't bother. Punctuality: Good speed"

S (1) = 0 S (2) =0.8 S (3) = 1

Calculating the total now using the above three calculated individual scores:

S = 1.63

After calculating the score for 500 posts for the above mentioned 4 trains, the mean score was calculated.

Depending upon the calculated means, Indian Railways can make the required improvements in appropriate trains and thus, this study can be extended to all the trains in the nation.

5. CONCLUSIONS AND FUTURE WORK

In this paper, the sentiment polarities of various feedbacks have been evaluated and thereby determining the general opinion on the experience of travelling in various trains of Indian Railways. We found that polarity dictionary generation enables to evaluate sentences in an effective manner and inclusion of quantifiers in the proposed method aptly reflects the intensities in various sentences. This research can be extended to various products and services. All the sentimental analysis problems are challenging because sentiments and emotions are extremely subjective but there's lot of scope of development and improvement in sentimental analysis. In future, we plan to extend our research to phrase level sentiment analysis i.e. there can be set of words which aren't polar words but still express an opinion. Also, we tend to address sentimental consistency and problems like capitalization and italicization of polar words which indicate different meaning as compared to regular usage of polar words. There are many opinions which use irony, satire, mockery, sarcasm etc. Apart from these sentimental analysis on exclamatory and negation sentences is also a new area. And ultimately it can be extrapolated to other languages as well, as the opinions are usually a mixture of two or more languages. Therefore, to exploit the complete benefits of sentimental analysis and natural language processing, sophisticated tools need to be developed which not only handle a large number of words but also various literary expressions.

6. ACKNOWLEDGEMENTS

We are indebted to the administrators of the blog www.indiarailinfo.com and also to all those people who take time out after their journeys to post their valuable feedbacks. This research would not have been possible without the data that we procured from this blog.

7. REFERENCES

[1] Ren, Fuji, Wu, Ye.2011. Learning Sentimental Influence on Twitter. International Conference on Future Computer Sciences and Applications.

[2] Mizumoto, Keisuke, Yanagimoto, Hidekazu, Yoshioka, Michifumi.2012. Sentimental Analysis of Stock Market News with Semi Supervised Learning. IEEE/ACIS 11th International Conference on Computer and Information Science

[3] Feifan, Liu, Dong,Wang, Li, Bin, Yang, Liu.2010. Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods. Human Language Technoligies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 309-312, Los Angeles, California, June 2010.

[4] Kaihui Zhang, Lei Li, Peng Li, Wenda Teng.2011. Stock Trend Forecasting Method Based on Sentiment Analysis and System Similarity Model. The 6th International Forum on Strategic Technology.

[4] Liu, Bing.2010. Sentimental Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F.J. Damerau), 2010

[5] Paltoglou, Georgios, Thelwall, Mike.2012. Twitter, MySpace, Dgg: Unsupervised Sentiment Analysis in Social Media. ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 4, Article 66, Publication date: September 2012.

[6] Taboada, Maite, Brooke, Julian, Tofiloski, Milan, Voll Kimberly, Stede, Manfred.2011. Lexicon-Based Methods for Sentiment Analysis. Association for Computational Linguistics. Volume 37, Number 2.

[7] Wilson, Theresa, Wiebe, Janyce, Paul Hoffmann.2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347-354, Vancouver, October 2005.

[8] Na, S.-H., Lee, Y., Nam, S.-H., and Lee, J.-H. 2009. Improving opinion retrieval based on query-specific sentiment lexicon. In ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, pages 734–738, Berlin, Heidelberg. Springer-Verlag.

[9] Ramanathan Narayanan, Bing Liu, Alok Choudhary.2009. Sentimental Analysis on conditional sentences. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 180–189, Singapore, 6-7 August 2009.

[10] Pang, L. Lee. and S. Vaithyanathan.2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP- 2002.

[11] Cynthia Whissell. 1989. The dictionary of affect in language.In R. Plutchik and H. Kellerman, editors, Emotion: theory, research and experience, volume 4. Acad. Press, London.

[12] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: Apublicly available lexical resource for opinion mining. In Proceedings of LREC 2006.

[13] Christiane Fellbaum. 1998 WordNet,an electronic lexical database. MIT Press.

[14] Beineke, Philip, Trevor Hastie, and Shivakumar Vaithyanathan. 2004. The sentimental factor: Improving review classification via human-provided information. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 263–270, Barcelona

[15] Das, Sanjiv Ranjan and Mike Y. Chen. 2001. Yahoo! for Amazon: Sentiment parsing from small talk on theWeb. In Proceedings of the August 2001 Meeting of the European Finance Association (EFA), Barcelona, Spain. Available at http://ssrn.com/ abstract=276189.