Context-based Sarcasm Detection in Hindi Tweets

Conference Paper · December 2017

DOI: 10.1109/ICAPR.2017.8593198

3 authors, including:

Drsantosh Kumar Bharti
Pandit Deen Dayal Petroleum University
42 PUBLICATIONS **517** CITATIONS

SEE PROFILE

Rahul Raman
VIT University
18 PUBLICATIONS **104** CITATIONS

SEE PROFILE

Context-based Sarcasm Detection in Hindi Tweets

1st Santosh Kumar Bharti
Dept. of CSE
NIT Rourkela
Rourkela, India
sbharti1984@gmail.com

1st Korra Sathya Babu
Dept. of CSE
NIT Rourkela
Rourkela, India
prof.ksb@gmail.com

2nd Rahul Raman
Dept. of CSE
VIT Vellore
Vellore, India
rahulraman2@gmail.com

***Abstract*—Sentiment analysis is the way of finding ones' opinion towards any specific target. Sarcasm is a special type of sentiment which infers the opposite meaning of what people convey in the text. It is often expressed using positive or intensified positive words. Nowadays, posting sarcastic messages on social media like Twitter, Facebook, WhatsApp, *etc.*, became a new trend to avoid direct negativity. In the presence of sarcasm, sentiment analysis on these social media texts became the most challenging task. Therefore, an automated system is required for sarcasm detector in textual data. Many researchers have proposed several sarcasm detection techniques to identify sarcastic text. These techniques are designed to detect sarcasm on the text scripted in English since it is the most popular language in social networking groups. However, parallel research for sarcasm detection on different Asian languages like Hindi, Telugu, Tamil, Urdu, and Bengali are not yet explored. One of the reasons for the less exploration of these languages for sarcastic sentiment analysis is the lack of annotated corpus even though they are popular in a large networked society. In this article, we proposed a context-based pattern *i.e.* "sarcasm as a contradiction between a tweet and the context of its related news" for sarcasm detection in Hindi tweets. The proposed approach utilized Hindi news as the context of a tweet with in the same timestamp and attained an accuracy of 87%.**

***Index Terms*—Hindi, Online News, Sarcastic, Sentiment, Social Media, Tweets**

I. Introduction

Online companion has gained tremendous momentum in recent times for business, politics, entertainment, *etc.* Social media such as Twitter, Facebook, WhatsApp, *etc.*, is considered as the popular medium for online companion and it attain the response of users from worldwide. These responses include ones' sentiment or opinion towards any specific target such as individuals, events, topics, products, organizations, services, *etc* [1]. The sentiment is nothing but an opinion of any individual towards a specific target. It may be either positive, negative or neutral. Manual extraction of the sentiment of the social media text is a tedious job for individuals as well as organizations. There is a need for an automated system which will be capable of providing sentiment of social media text without any human interference.

Sentiment analysis is a part of Natural Language Processing (NLP) that deals in finding the orientation of an opinion in a piece of text about any topic [2]. The presence of sarcastic text in the corpus makes sentiment analysis challenging as most of the existing sentiment analysis system does not consider sarcastic sentiments. Due to this, most of the existing systems for sentiment analysis fail in detecting the sarcastic sentiment. Sarcasm is a special kind of sentiment that usually flips the orientation of the opinion in a given piece of text. People often express sarcasm verbally through the use of heavy tonal stress and certain gestural clues like rolling of the eyes, hand's movement, *etc.* These tonal and gestural clues are obviously missing while expressing sarcasm in text, which makes its detection even more difficult task. Sometimes, human beings feel difficulty to understand sarcasm in text. The sarcastic sentence usually looks positive, but overall meaning becomes negative due to the presence of sarcasm. An automated system is required for sentiment analysis which will be capable of identifying sarcastic sentiment as well.

In recent past, many researchers have focused on sarcasm detection and proposed automatic sarcasm detector in text [3]–[12]. These systems are developed for sarcasm detection in text scripted in English. English is the most popular language across the world, and in this domain, plenty of resources are freely available for research. Therefore, majority of researchers have preferred English domain for their research. Many other languages are getting popular in rapid pace such as Hindi, Arabic, Dutch, Mandarin, etc. These languages fall in low resourced categories as the availability of free resources in this domains are very rare.

In low resourced languages, Hindi is the fourth-most spoken language in the world, after Mandarin, Spanish and English [13]. It has 490 million speakers across the world, and majority of them are from India [14]. It is widely used for speaking in countries like India, Mauritius, Fiji, Suriname, Guyana, Trinidad & Tobago and Nepal [15]. These days' in India, Hindi is getting more popularity on social media such as Facebook, Twitter, WhatsApp, etc. People are posting messages, comments very frequently in the Hindi language. With the increased amount of information being communicated via regional languages like Hindi on social media, there comes a promising opportunity of mining this information. In order to mine the Hindi information automatically from social media, various NLP tasks such as part-of-speech (POS) tagging, sentiment analysis had been already developed.

For sarcasm detection in Hindi, a system was developed [16] for Hindi tweets using a similar set of features used for English tweets namely, #tag, emoticons, punctuation marks

The value of precision and recall may vary application to application. For example, an application attains high precision but low recall. Similarly, another application attains low precision but high recall. To deal with this situation, one can rely on another statistical parameter *i.e.* $F1\text{-}measure$. It is a harmonic mean of $precision$ and $recall$. To obtain the $F1\text{-}measure$, a formula is given in Equation 5 .

$$F1\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

A set of 1000 random Hindi tweets is used as a testing set to experiment. After the experiment, a confusion matrix of 1000 Hindi tweets is given in Table III for error analysis. To identify sarcasm in tweets, the proposed system identified 348 tweets correctly out of 370 sarcastic tweets. The ground truth of sarcastic and non-sarcastic tweets are given in Table II.

TABLE III
CONFUSION MATRIX FOR SARCASM DETECTION IN HINDI TWEETS.

Proposed approach	No. of tweets	T_p	T_n	F_p	F_n
Identifying sarcasm	1000	348	522	62	68

Using the confusion matrix given in Table III, the values of $precision$, $recall$, $F1\text{-}measure$ and $accuracy$, attained by the proposed approach for identifying sarcasm in tweets are given in Table IV. Finally, we made a comparison of the proposed approach with state-of-the-art approaches and is shown in Table V.

TABLE IV
ACCURACY, PRECISION, RECALL AND F1-MEASURE

Proposed approach	$Precision$	$Recall$	$F1\text{-}score$	$Accuracy$
Identifying sarcasm	0.848	0.836	0.842	0.87

TABLE V
COMPARISON OF PROPOSED APPROACH WITH SOME OF THE STATE-OF-THE-ART TECHNIQUES FOR SARCASM DETECTION IN HINDI TWEETS.

Study	$Precision$	$Recall$	$F1\text{-}score$	$Accuracy$
Desai *et al.*	0.732	0.674	0.705	0.714
Bharti *et al.*	0.736	0.717	0.726	0.794
Proposed Approach	**0.848**	**0.836**	**0.842**	**0.87**

VI. CONCLUSION

In the absence of sufficient dataset for training and testing, detection of sarcastic sentiment is a challenging task in Hindi. This article proposed a context-based pattern for sarcasm detection in Hindi tweets. Online News from Twitter news sources is utilized as the context of the tweets in the same timestamp. The proposed approach attains an accuracy of 87%. The proposed approach outperforms the state-of-the-arts techniques for sarcasm detection in Hindi tweets.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of conference on Empirical methods in natural language processing*, vol. 10, no. 1. ACL, 2002, pp. 79–86.

[3] C. Liebrecht, F. Kunneman, and A. van den Bosch, "The perfect solution for detecting sarcasm in tweets# not," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. New Brunswick, NJ: ACL, 2013, pp. 29–37.

[4] R. González-Ibánez, S. Muresan, and N. Wacholder, "Identifying sarcasm in twitter: a closer look," in *Proceedings of the 49th Annual Meeting on Human Language Technologies*. ACL, 2011, pp. 581–586.

[5] F. Barbieri, H. Saggion, and F. Ronzano, "Modelling sarcasm in twitter a novel approach," *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 50–58, 2014.

[6] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proceedings of the conference on empirical methods in natural language processing*, 2013, pp. 704–714.

[7] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in twitter data," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ACM, 2015, pp. 1373–1380.

[8] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcastic sentences in twitter and amazon," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. ACL, 2010, pp. 107–116.

[9] F. Kunneman, C. Liebrecht, M. van Mulken, and A. van den Bosch, "Signaling sarcasm: From hyperbole to hashtag," *Information Processing & Management*, vol. 51, no. 4, pp. 500–509, 2015.

[10] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on twitter: A behavioral modeling approach," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 97–106.

[11] S. Bharti, B. Vachha, R. Pradhan, K. Babu, and S. Jena, "Sarcastic sentiment detection in tweets streamed in real time: a big data approach," *Digital Communications and Networks*, vol. 2, no. 3, pp. 108–121, 2016.

[12] S. K. Bharti, R. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection based on types of sarcasm occurring in twitter data," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 13, no. 4, pp. 89–108, 2017.

[13] M. Parkvall, "Världens 100 största språk 2007," *The World's*, vol. 100, 2007.

[14] Language and Culture, "Top 30 languages by number of native speakers," 2015. [Online]. Available: http://www.vistawide.com/languages/top-30-llanguages.htm/

[15] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of Language Resources and Evaluation Conference*, 2006, pp. 417–422.

[16] N. Desai and A. D. Dave, "Sarcasm detection in hindi sentences using support vector machine," *International Journal*, vol. 4, no. 7, pp. 8–15, 2016.

[17] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2013, pp. 195–198.

[18] P. Liu, W. Chen, G. Ou, T. Wang, D. Yang, and K. Lei, "Sarcasm detection in social media based on imbalanced classification," in *Web-Age Information Management*, 2014, pp. 459–471.

[19] S. Bharti, S. Korra, and S. Jena, "Harnessing online news for sarcasm detection in hindi tweets." *Proceeding of PReMI 2017 to be published by LNCS*, 2017.

[20] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 133–140.

[21] A. Bharati, R. Sangal, D. M. Sharma, and L. Bai, "Anncorra: Annotating corpora guidelines for pos and chunk annotation for indian languages," *LTRC-TR31*, 2006.

[22] J. L. Fleiss, J. Cohen, and B. Everitt, "Large sample standard errors of kappa and weighted kappa." *Psychological Bulletin*, vol. 72, no. 5, p. 323, 1969.