Harnessing Online News for Sarcasm Detection in Hindi Tweets

Santosh Kumar Bharti (orcid.org/0000-0002-0627-6433), Korra Sathya Babu
(orcid.org/0000-0002-5963-5735), Sanjay Kumar Jena
(orcid.org/0000-0002-1619-8360)

Department of Computer Science and Engineering, National Institute of Technology, Rourkela,
Odisha, India.

Abstract. Detection of sarcasm in Indian languages is one of the most challenging tasks of Natural Language Processing (NLP) because Indian languages are ambiguous in nature and rich in morphology. Though Hindi is the fourth popular language in the world, sarcasm detection in it remains unexplored. One of the reasons is the lack of annotated resources. In the absence of sufficient resources, processing the NLP tasks such as POS tagging, sentiment analysis, text mining, sarcasm detection, *etc.*, becomes tough for researchers. Here, we proposed a framework for sarcasm detection in Hindi tweets using online news. In this article, the online news is considered as the context of a given tweet during the detection of sarcasm. The proposed framework attains an accuracy of 79.4%.

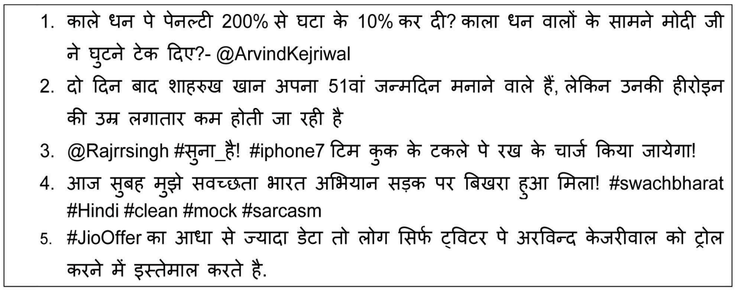
Keywords: Hindi tweets, NLP, Online news, Sarcasm, Sentiment.

1 Introduction

With 490 million speakers [1] across the world, Hindi stands fourth in popularity after Mandarin, Spanish, and English [2]. In social media such as Twitter, Facebook, WhatsApp, *etc.*, most of the Indians now prefer Hindi for communication, and this generates large volumes of data. The manual process of mining the sentiments from these large data is a tedious job for individuals as well as organizations. Therefore, an automated system is required to identify the sentiment automatically from Hindi text.

Sentiment analysis is a task which identifies the orientation of a text towards a specific target such as products, individuals, organizations, *etc.* With the presence of sarcasm, the prediction of sentiment in text often goes wrong in the analysis. Sarcasm often conveys negative meaning using positive or intensified positive words. For example, “I love waiting forever for the doctor”. In the first look, the sentence conveys positive sentiment; but, it is sarcastic. Due to this, most of the existing sentiment analyzers fail to detect real sentiment.

Recently, many sarcasm detectors were developed by researchers for text scripted in English[3-9]. But, there is only one reported work available for detection of sarcasm in Hindi scripted text [10]. The existing work [10] does not consider the natural Hindi tweets [1] for the experiment. Their training and testing set consists of Hindi tweets translated from English scripted tweets. In this article, we proposed a framework for sarcasm detection in natural Hindi tweets using online Hindi news as the context. A sample of natural Hindi sarcastic tweets is shown in Fig. 1.



1. काले धन पे पेनल्टी 200% से घटा के 10% कर दी? काला धन वालों के सामने मोदी जी ने घुटने टेक दिए?- @ArvindKejriwal
2. दो दिन बाद शाहरुख खान अपना 51वां जन्मदिन मनाने वाले हैं, लेकिन उनकी हीरोइन की उम्र लगातार कम होती जा रही है
3. @Rajrrsingh #सुना_है! #iphone7 टिम कुक के टकले पे रख के चार्ज किया जायेगा!
4. आज सुबह मुझे सवच्छता भारत अभियान सड़क पर बिखरा हुआ मिला! #swachbharat #Hindi #clean #mock #sarcasm
5. #JioOffer का आधा से ज्यादा डेटा तो लोग सिर्फ ट्विटर पे अरविन्द केजरीवाल को ट्रोल करने में इस्तेमाल करते है.

Fig. 1: A sample of Hindi sarcastic tweets.

Tweets and news are very similar in nature as both describes current happenings in their way. The news gives us the authenticated knowledge about real-time happenings across the world. Similarly, users‘ from worldwide shares their feeling on current happening through tweets. It may or may not be authentic. It depends on the individual user and their likes and dislikes. If a user likes any current happenings, then they will share positive feeling on that happenings. If they do not like, then they may share either direct negative or sarcastic feeling. In this approach, news has been utilized as the context of the given tweet to predict the authenticity of the tweet with the truth. If a given tweet follows the orientation of the related news, then is be considered as a simple tweet, and the obtained sentiment is correct. If the tweet does not follow the orientation of the related news, then the tweet is classified as sarcastic, and the obtained sentiment is opposite.

The rest of the paper is organized as follows: Section 2 describes related work. The proposed scheme is discussed in Section 3. Analysis of the results are given in Section 4 and conclusion of the article is drawn in Section 5.

2 Related Work

Sarcasm detection in resource rich language like English is well explored [3-9]. In the context of Indian languages, it is yet to be explored. The main reason is the unavailability of benchmark resources for training and testing.

Desai *et al.* [10] proposed a Support Vector Machine (SVM) based sarcasm detector for Hindi sentences. They used Hindi tweets as the dataset for training and testing using SVM classifier. In the absence of annotated datasets for training and testing, they converted English tweets into Hindi. Therefore, they focused on a similar set of features like emoticons and punctuation marks for sarcasm detection in English text. These methods are not applied directly for the natural Hindi sarcastic tweets as shown in Fig. 1.

[1] A natural Hindi tweet is a tweet that is available on Twitter in natural Hindi language unlike translated from English to Hindi.

3 Proposed Scheme

This section describes the proposed framework for sarcasm detection in Hindi tweets as shown in Fig. 2. Here, online news is used as a context which authenticates the given tweets with actual happenings. Here, we assume that online news is correct and authenticated.

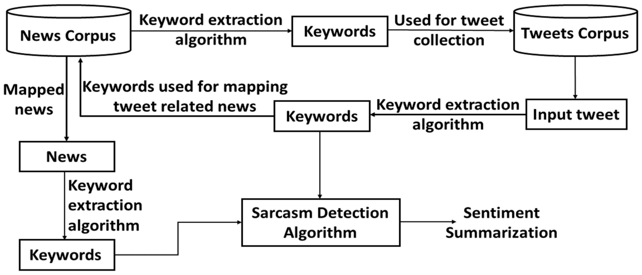

News Corpus
Keyword extraction algorithm
Keywords
Used for tweet collection
Tweets Corpus
Mapped news
Keywords used for mapping tweet related news
Keywords
Keyword extraction algorithm
Input tweet
News
Keyword extraction algorithm
Sarcasm Detection Algorithm
Sentiment Summarization
Keywords

Fig. 2: Proposed framework for sarcasm detection in Hindi tweets.

For every news in the authenticated news corpus, keywords are extracted using Algorithm 1. These keywords are used to obtain the possible tweets. Further, for prediction of a sarcastic tweet, it takes a tweet as an input and extracts the important keywords using Algorithm 1. Then, the extracted keywords are used to map the related authenticated news in news corpus. Finally, it fed both the sets of keywords (input tweet and related news) to sarcasm detection algorithm to classify the tweet is sarcastic or not.

3.1 News Collection

After browsing several online news sources, we have collected a total of around 5000 one liner Hindi news manually on recent topics from top rated news sources as mentioned in Fig. 3. The collected news belongs to different categories such as sports, movies, business, politics, *etc*. In the preprocessing, redundant news are eliminated. News related to murder, rape, bomb blast, *etc*. were discarded. We believe that sarcastic tweets will not be floated on serious topics. It was thus eliminated. After preprocessing, the news corpus consists of a total of 2000 authenticated unique news.

3.2 Keyword Extraction

This section describes the procedure of keyword extraction from sentences as shown in Algorithm 1. Algorithm 1 takes authenticated news corpus $(\mathtt{C})$ as an input and find Part-of-Speech (POS) tag information for every news in the corpus. For every news, the tags noun (NN), verb (V), adjective (ADJ) and adverb (ADV) are extracted from the tagged set, and the corresponding tokens are extracted as $\langle Set\ of\ Keywords \rangle$ for that news.

POS Tagging To identify the POS tag information in Hindi sentences, we have developed a Hidden Markov Model (HMM) based POS tagger. It uses Indian Language (IL) standard tagset which consists of 24 tags [11]. For example, the POS tag information

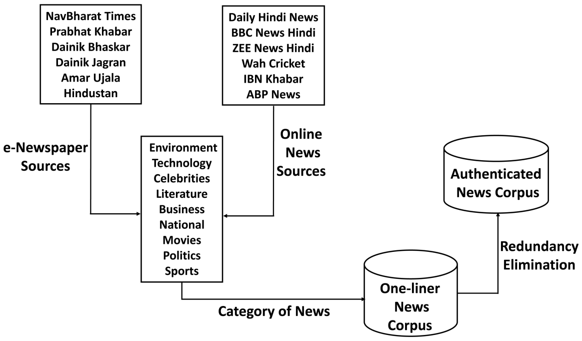


Fig. 3: Procedure for news collection.

Algorithm 1: Keywords_Extraction_Algorithm

```
Data: $dataset$ := Corpus of authenticated news ($\mathbb{C}$)
Result: $classification$ := $\langle Set\ of\ Keywords \rangle$ for every news in the corpus
Notation: $ADJ$: Adjective, $V$: Verb, $ADV$: Adverb, $NN$: Noun, $NS$: News sentence,
  $\mathbb{C}$: Corpus, $T$: Tag, $K$: Keyword, $NTS$: News-wise tagged set, $NKS$: News-wise set of
  keywords, $LoK$: List of Keywords.
Initialization : $NKS$ = { $\phi$ }, $LoK$ = { $\phi$ }
while $NS\ in\ \mathbb{C}$ do
    $NTS$ = find_POS_tag ($NS$)
    while $T\ in\ NTS$ do
        if $(T == (ADJ||V||ADV||NN))$ then
            $K \leftarrow$ Keyword[$T$]
        end
        $\langle NKS \rangle \leftarrow NKS \cup K$
    end
    $LoK \leftarrow LoK \cup \langle NKS \rangle$
end
```

of Hindi sentence क्या आपका नाम राम है? is क्या - WQ | आपका - PRP | नाम - NN | राम - NNP | है - VAUX | ? - SYM |. The Hindi POS tagger tool is available on URL:http//www.taghindi.herokuapp.com.

3.3 Tweets Collection

To get the news related tweets, we used extracted $\langle Set\ of\ keywords \rangle$ for every news from news corpus to collect the possible tweets from Twitter as shown in Fig. 4. On deploying all the sets of keywords from 2000 unique news, a total of around 5000 Hindi tweets is collected. A sample set of news and related tweets are released on URL: https://github.com/rkp768/hindi-pos-tagger/tree/master/
News%20and%20tweets.

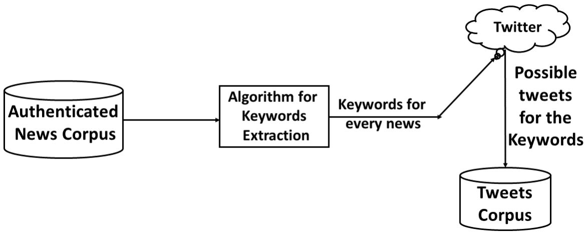

Twitter
Authenticated News Corpus
Algorithm for Keywords Extraction
Keywords for every news
Possible tweets for the Keywords
Tweets Corpus

Fig. 4: Procedure of tweets collection.

3.4 Sarcasm Detection

In this section, an algorithm is proposed to classify the tweet as sarcastic or not in the context of online news information. The procedure of identifying sarcastic tweet is given in Algorithm 2.

```
Algorithm 2: Sarcasm_Detection_using_Online_News
Data: $dataset$ := $\langle Set\ of\ Keywords \rangle$ for both input tweet and corresponding related
      news.
Result: $classification$ := Input tweet is sarcastic or not.
Notation: $\langle SoK \rangle_n$: Set of Keywords for news, $\langle SoK \rangle_t$: Set of Keywords for input tweet,
  $(PKC)_n$: positive keywords in news, $(PKC)_t$: positive keywords in input tweet,
  $(NKC)_n$: negative keywords in news, $(NKC)_t$: negative keywords in input tweet
if ($\langle SoK \rangle_n == \langle SoK \rangle_t$) then
    Tweet is not sarcastic.
end
else
    $(PKC)_n$ = Count_postive_keywords($\langle SoK \rangle_n$)
    $(NKC)_n$ = Count_negative_keywords($\langle SoK \rangle_n$)
    $(PKC)_t$ = Count_postive_keywords($\langle SoK \rangle_t$)
    $(NKC)_t$ = Count_negative_keywords($\langle SoK \rangle_t$)
end
if $(PKC)_n > (PKC)_t$ then
    Tweet is sarcastic.
end
else if $(NKC)_n > (NKC)_t$ then
    Tweet is sarcastic.
end
else
    Tweet is not sarcastic.
end
```

The Algorithm 2 takes both the sets of keywords (one for input tweet and other for related news) as the input. Then, it compares both the sets of keywords. If both the sets contain similar keywords, it means the orientation of the news and tweet are same. Therefore, the tweet is authentic and not sarcastic. If both sets do not contain similar keywords, then it calculates the number of positive and negative keywords in both news and tweet using a predefined list of Hindi words with polarity value. The list of Hindi

SentiWordNet is available on URL:https://github.com/smadha/SarcasmDetector/blob/master/Hindi%20SentiWordNet/HSWN_WN.txt. Further, it compares the count of positive and negative keywords. If the news contains more positive keywords than an input tweet, it indicates the user intentionally negate the temporal fact (news). In this case, the orientation of the news is positive, and the orientation of the tweet is negative. Due to this contradiction, given input tweet is classified as sarcastic. Similarly, in the case of more negative keywords in the news than input tweet, given tweet is classified as sarcastic. For rest of the cases, tweets are not sarcastic.

4 Results and Discussion

This section describes the experimental results of the proposed approach to identify sarcasm in Hindi tweets. To test the performance, four experimental parameters have been used namely, $Precision$, $Recall$, $F1$-$measure$ and $Accuracy$. A set of 500 random tweets from collected Hindi tweets corpus is used as a testing set to experiment. To annotate the testing set as sarcastic or not, three annotators are used, and the results of annotators are used as ground truth while testing. A confusion matrix for identifying sarcasm in 500 tweets are given in Table 1. Using the confusion matrix given in Table 1, the values of precision, recall, F1-measure and accuracy attained by the proposed approach for identifying sarcasm in Hindi tweets are given in Table 2.

Table 1: Confusion matrix for sarcasm detection in Hindi tweets.

Proposed approach	No. of tweets	T_p	T_n	F_p	F_n
Identifying sarcasm	500	137	260	51	56

Table 2: $Precision$, $Recall$, $F1$-$measure$ and $Accuracy$ attained by proposed approach

Proposed approach	$Precision$	$Recall$	$F1$-$measure$	$Accuracy(\%)$
Identifying sarcasm	0.736	0.717	0.726	79.4

While identifying sarcasm in Hindi tweets concerning news context, we consider the comparison of $\langle Set\ of\ Keywords\rangle$ for both input tweet and corresponding related news. We assume all the news have neutral sentiments whereas tweets contain either positive, negative or neutral sentiment. Therefore, instead of sentiment comparison, we preferred the comparison of individual keywords and its orientation. If both news and tweets describe same orientation, then the tweet is non-sarcastic. If the orientation of news and tweet are not same, it means the user is trying to negate this temporal fact intentionally. Hence, the given input tweet is sarcastic.

Limitations The proposed framework has the following limitations:

1. In this research, news time-stamp is not available. Hence, while mapping a tweet to a unique related news, we are fully dependent on keywords, which does not give full assurance that the news and tweet belong to the same time-stamp.
2. If few keywords are matched for news and tweet, but both belong to different time-stamp. In such situation prediction of sarcasm may or may not be correct.

5 Conclusion and Future Direction

In the absence of sufficient annotated dataset for training and testing, one can not apply traditional methods for sarcasm detection in Hindi tweets that are used in examples. Therefore, this article proposes a novel framework for sarcasm detection in Hindi tweets using the online news as context. As news usually carry neutral sentiment, we used the important keywords for both input tweet and its related news to decide the tweet is sarcastic or not concerning the related news. The proposed approach attains 79.4% accuracy.

In future, we will resolve the current limitation of the article. The framework will be updated with time-stamp verification while mapping a tweet to the news.

References

1. M. Parkvall.: Varldens 100 storsta sprak 2007. The Worlds 100, 2007.
2. W. Language and Culture.: Top 30 Languages by Number of Native Speakers, 2005. http://www.vistawide.com/languages/top 30 languages.htm
3. C. Liebrecht, F. Kunneman, and A. van den Bosch.: The perfect solution for detecting sarcasm in tweets# not. In proceedings of the Association for Computational Linguistics, pp. 29-37, 2013.
4. R. Gonzalez-Ibanez, S. Muresan, and N. Wacholder.: Identifying sarcasm in twitter: a closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, (2), pp. 581-586, 2011.
5. A. Joshi, V. Sharma, and P. Bhattacharyya.: Harnessing Context Incongruity for Sarcasm Detection. In ACL (2), pp. 757-762, 2015.
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 Empirical methods in natural language processing, pp. 704-714, 2013.
7. S. Bharti, K. Sathya Babu, and S. Jena.: Parsing-based sarcasm sentiment recognition in twitter data. In International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1373-1380, IEEE/ACM, 2015.
8. S.K. Bharti, B. Vachha, R.K. Pradhan, K.S. Babu, and S.K. Jena.: Sarcastic sentiment detection in tweets streamed in real time: a big data approach. Digital Communications and Networks 2(3), 108-121, 2016.
9. S.K. Bharti, R.K. Pradhan, K.S. Babu, and S.K. Jena.: Sarcasm Analysis on Twitter Data Using Machine Learning Approaches. Trends in Social Network Analysis. Springer International Publishing, 51-76, 2017.
10. N. Desai, A.D. Dave.: Sarcasm Detection in Hindi sentences using Support Vector machine. International Journal 4(7), 815, 2016.
11. A. Bharati, R. Sangal, D. Sharma, D. and L. Bai.: Anncorra: Annotating corpora guidelines for pos and chunk annotation for indian languages. LTRC-TR31, 2006.