e-ISJN: A4372-3114
p-ISJN: A4372-3115

ISSN: 2321-7782 (Online)
ISSN: 2347-1778 (Print)
Impact Factor: 7.327

Volume 7, Issue 8, August 2019

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

A Comprehensive Study of Challenges, Opportunities in Sarcasm Detection on Textual Data

Anandkumar D. Dave
Department of Information Technology,
Dharmsinh Desai University,
Nadiad – India.

Abstract: In recent years huge amount of textual data is being generated on the web and this allowed researchers to explore new dimensions of mining the meaningful information from it. Although Sentiment Analysis is one of the trending areas, and it is widely explored, sarcasm plays vital role in influencing the results of sentiment analysis. According to cambridge dictionaries, the sarcasm is defined as "the use of remarks that clearly means the opposite of what they say, or criticize something in a humorous way". Moreover, growth of translation industries [31] has allowed diversity of languages in this enormous textual data. In this paper will highlight key challenges, problems and opportunities available, in order to detect the sarcasm in textual sentences.

Keywords: Sentiment Analysis, Opinion Mining, Sarcasm Detection, Natural Language Processing, Text Processing, Machine Learning.

I. INTRODUCTION

In context of processing the textual data for sentiment analysis has its own issue to deal with. Some of these issues are sarcasm detection, thwarted expressions, unstructured data, noise, contextual information and Word Sense Disambiguation. Among these sarcasms directly affects the results of sentiment analysis as it changes the sentiment or polarity of target sentence if not detected and corrected [28]. Due to advancement in technology and easy availability, and awareness among the users, enormous amount of data in diverse languages is generated electronically in form of opinions, feedbacks, and emotional view over various topics on different platform such as Twitter, Facebook, On line retail portals. These unstructured data is predicted to grow up to 40 zettabytes and translation industries will grow up to $30 billion business by 2020[30]. At current stage, more than 12 Indian languages are on list of top 30 languages spoken by the native speaker over the world [20]. A survey [31] shows that 70% - 80% of buyers are influenced by the other's opinions or reviews on specific product. So, In near future this phenomenon will need a state of art solutions for processing electronic data. So, necessity of development of various Natural Language Processing tool and Machine Learning technological support for diversified languages is must, and this will pose new opportunities among the research community along with its own challenges.

Sarcasm is the part of British culture [32]. Sarcasm is characterized as the ironic or satirical wit that is intended to insult, mock, or amuse [19]. Sarcasm is an elegant way of the conveying message so that it is hard to detect. Such sarcastic sentences may impose problems for many Natural Language Processing and Machine Learning based system such as online review summarization systems, dialogue systems or brand monitoring systems. Moreover, linguistics diversity makes it more tedious problem due to lack of availability of tools and techniques.

This paper is focused on challenges, opportunities available in detection of sarcastic sentences presented in the textual form like "I love being ignored".

II. Related Work

Researchers have considered sarcasm detection as a challenging task in context of text mining [16]. Today,researchers are considering text mining problems as a bigdata problems due to extreme amount of the data availability. Sarcasm Detection is aimed to find out correct emotions, feelings, or attitude of writer from given text segment [28]. Today, online retailers, social media handlers have abandon amount of textual data, that can be processed to gain some meaningful hidden information or pattern that can be utilized by various industries. Sarcasm detection in textual sentences inherits the fundamental challenges from Natural Language Processing such as Lexical, structural, semantic and pragmatic ambiguities. These challenges have become more heftier due tomultilingual nature of data. However, sarcasm in speech can beidentified by tone, context, gesture-posture and some of pausesin the sentences. This also requires good hold over the language, sometime even expert users of language cannot identify the spoken sarcasm and it became worse while it comes in textual form. Sarcasm is studied well by psychologists, behavioral scientists and linguistics [20]. Pang and Lee, (2008) have explored the area of opinion mining in textual form. While spoken form of sarcasm was studied by Tepperman et al. (2006) with the use of 'yeah-right' pronunciation. Kreuz and Caucci (2007) studied the prejudicedeffect of different lexical factors like interjections and punctuation symbols in recognizing sarcasm in written form. Annotated corpus for sarcasm detection of Amazon product reviews had highlighted by Filatova(2012). Tsur et al.'s (2010)framework has used syntactic and pattern based features to detect the sarcasm in Amazon product reviews. Moreover, Devidov et al. (2010) have used sarcastic Twitter message and Amazon product review to train the classifier using syntactic and pattern based feature [18] and they have considered 'hashtag' as promising marker to annotate the corpora but theirexperiments concludes that 'hashtags' are noisy. For this research, 75000 tweets were collected by [19] assuming that human tags are correct. Then a machine learning classifier is trained then applied it to test set of 3.3 million Dutch tweets. Classifier has correctly classified 101 tweets out of 135. Classifier is tested on top 250 tweets which are ranked, likely to be sarcastic, and achieved only 30% precision on it. The result shows that it is hard to detect sarcasm from literal tweets. Results have shown that sarcasm is often detected by hyperbole, using intensifier and exclamation. Non-hyperbolic sarcastic message needs external marking. In order to detect sarcasm authors [21] have defined sarcastic statement as positive sentiment with negative situation. Eg. I love, being ignored. Micro blogging web site like twitter they have found this common structure of sarcastic tweet. They made assumption that a positive sentiment verb phrase usually appears to the left of a negative situation phrase and in close proximity (usually, but not always, adjacent). Pictorially, we assume that many sarcastic tweets contain this structure:

III. Challenges for Sarcasm Detection

Researchers have considered sarcasm detection as a challenging task in context of text mining [16]. Today,researchers are considering text mining problems as a bigdata problems due to extreme amount of the data availability. Sarcasm Detection is aimed to find out correct emotions, feelings, or attitude of writer from given text segment [28]. Today, online retailers, social media handlers have abandon amount of textual data, that can be processed to gain some meaningful hidden information or pattern that can be utilized by various industries. Sarcasm detection in textual sentences inherits the fundamental challenges from Natural Language Processing such as Lexical, structural, semantic and pragmatic ambiguities. These challenges have become moreheftier due to multilingual nature of data. However, sarcasm inspeech can be identified by tone, context, gesture-posture and some of pauses in the sentences. This also requires good hold over the language, sometime even expert users of language cannot identify the spoken sarcasm and it became worse while it comes in textual form. Sarcasm is studied well by psychologists, behavioral scientists and linguistics [20]. Pang and Lee, (2008) have explored the area of opinion mining in textual form. While spoken form of sarcasm was studied by Tepperman et al. (2006) with the use of 'yeah-right' pronunciation. Kreuz and Caucci (2007) studied the prejudiced effect of different lexical factors like interjections and punctuation symbols in recognizing sarcasm in written form. Annotated corpus for sarcasm detection of Amazon product reviews had highlighted by Filatova(2012). Tsur et al.'s (2010) framework has used syntactic and pattern-based features to detect the sarcasm in Amazon product reviews. Moreover, Devidov et al. (2010) have used sarcastic Twitter message and Amazon product review to train the classifier using syntactic and pattern-based feature [18] and they have considered 'hashtag' as promising marker to annotate the corpora but their experiments conclude that 'hashtags' are noisy.

IV. Current Limitations

Fundamental issues of NLP like Morphological analysis, Part of Speech tagging, chunking, Parsing, Semantic of Conceptualization, are the major barrier for sarcasm detection. Today, majority of reviews, feelings and opinions are available in English language and majority of work has been done for English language. So, the support of tools is mainly developed for English language only and have gain named Hindi-SentiWordNet (H-SWN). Their work is the first efforts to make rich lexical resources from well-developed languages such as English to Hindi and their results shows in- language sentiment analysis is having upper hand. Sarcasm detection in different language needs language specific tools, standard dictionaries and domain specific corpora. Santosh Kumar Bharti et.al[34] has emphasize on necessity of annotated corpora. Due to the lack of for-said resources, they introduced keyword-based algorithms to detect the sarcastic in online news collected from various online platforms. Ebrahimi M, Yazdavar AH, Sheth A had studied the effect of highly dynamic events on sentiment analysis. They collected 10,000 tweets for five candidates stood in presidential election in July 2016 in US. After their results they also conclude to include positive and negative words as we [33] used in our previous work. They also find data set generation as a fundamental challenge in sentiment analysis also these events are highly dynamic in nature. This will affect the data set that was build few weeks back and may results in appropriate results. The table 1.1 will show the statistic of various work done in sarcasm detection in various languages by researchers. significant success. However, growth of translation industries, have enabled individuals to express their views in their own languages. Subsequently, researches have also explored the Dutch and Indonesian languages for sarcasm detection [17][23]. Work carried out by Aditya Joshi, Balamurali A R and Pushpak Bhattacharyya for sentiment analysis in Hindi language [2] have come with standard corpora of Movie review in Hindi language, and developed lexical resource.

TABLE 1.1 COMPARISON OF WORKS

Index	Domain Dependent	Data Set	Classification Techniques		Measures *Accuracy*	Reference	Language	Sentence Length
1	Yes	Twitter and Amazon	Corpus Based	K-nearest neighbor	82	[18]	English	Not Given
2	Yes	Twitter	Corpus Based	SVM with sequential minima Optimization and Logistic Regression	71 66	[16]	English	Not Given Not Given
3	Yes	Twitter	Corpus Based	Winnow Classifier	79	[17]	Dutch	Not Given
5	Yes	Twitter	Corpus Based	Decision Tree, SVM and Logistic Regression	83.05 83.46 78.06 83.46	[20]	English	Not Given Not Given Not Given Not Given
6	Yes	Twitter	Corpus Based	Binary Logistic Regression	84.03	[21]	English	Not Given
7	Yes	Social Media Messages	Corpus Based	SVM, Naïve Bayed and Maximum Entropy	53.10 54.1 53.8	[23]	Indonesian	Not Given Not Given Not Given
8	Yes	Travell Review	Corpus Based	SVM	83		Marathi and Hindi	4-5 sentence. 10 words each

V. Experiment and Results

In our past work [33] we have tried to detect the sarcasm in Hindi sentences collected from various online resources. We have 1410 sentences with us and we also prepared Word-Antonym Pair, Positive and Negative word lists, Cue words list, intensity-based list of emoticons as marker to detect the sarcastic sentences with well-defined approach. In this work is the extension of previous work and more focused on length of the sentence in terms of words per sentences and number of characters per word and found that longer the sentenced has less chances of being detected as sarcastic sentence. For an example, इस f

ĭ ो dद] हdगfहY1. यह Y हd f ĭ त]ह लग]ह ह1, Jfåd`त

हलk ` ह, औ] Jfåd`त द åी Jŭ ा ह, f] åी ह द

ो 5 dहk ा 1 हल f इस ` दे ा] स` Jf द` ग |

Above sentence has more Positive- Negative words, noemoticons, and no Cue words. However, the word "हलांकि" is used as transition word which is changing the meaning of sentence and will result in sarcastic sentence by manual annotation. But will be difficult to detect automatically.

Moreover, the word " dद] हd गfहY" has context dependent meaning and hard to detect with limited annotated corpora. So, such language specific issues such as these transition words and context dependent meaning to assign proper polarity to sentence to have well efficient process to detect the sarcasm. In this approach, the experiments were performed in three setups as explained below. 10-fold cross validation is applied in each setup. Accuracy is the parameter for evaluation for all the experiments as well for comparison of results. The features are Unigram (TFIDF), Pos. Score , Neg. Score, #tag , Emoticons and Polarity values.

1) Experiment setup I: SVM classifier is trained with sentences which includes all markers available in those sentences.

   Total sentences: - 1410 sentences (of Variable length of wordsand characters per sentences)

   Classifier settings: - Support Vector Machine Kernel: - linearTarget classes: - two (sarcastic, non-sarcastic)

   Accuracy Achieved: 65 %

2) Experiment setup II: SVM classifier is trained with sentences which includes all markers available in those sentences.

   Total sentences in corpus: 750 sentences (Limited length ofwords and characters or one-line sentences includes markers) Model settings: - LibSVM

   Kernel:- C-SVM, Type: RBF, C=0.5, gamma = 0.007

   Target classes: - five (extreme positive /extreme negative/mildpositive /mild negative sarcastic/ non sarcastic).

   Accuracy achieved: 83%

   Experiment setup III: Data set with no external markers like #Tag, Emoticons or "!" marks. For such longer sentences. (Sentences with the conjunctions and complex sentences more than 20 words.)

3) Model settings: - LibSVM
4) Kernel: - C-SVM, Type: RBF, C=0.3, gamma = 0.007 Target classes: - two (sarcastic, non- sarcastic) Accuracy achieved: - 40%
5) For all the previous experiment setting we try to increase the length of sentences the result decreased drastically upto 20%-30% as sentences are over getting overlap with the various other challenges of the Challenges of Sentiment Analysis such as Identifying subjective portions of text, Domain dependence, Thwarted expressions [highly related with the length of the sentences.], Indirect negation of sentiment, Order dependence [In Text Classification structure does not play any role in classification because words are considered independent of each other].

VI. Conclusion

Automatic sarcasm detection is challenging issue and availability of language specific resources play the vital role in detection of sarcastic sentences in particular language. Sarcasmdetection in Hindi, Dutch, and Bangla languages has been tried and results has shown the fact that the annotated corpora has played significant role in training the machine learning algorithm to detect the sarcastic sentences. However, fromresults, it can be concluded that not only annotated corpora but the length of the sentences are also playing the vital role. As the over lapping context such as Domain dependence, Thwarted expressions [highly related with the length of the sentences.], Indirect negation of sentiment, Order dependence [In Text Classification structure does not play any role in classification because words are considered independent of each other] are become the main causes of the classifying the sarcastic sentences in to the non-sarcastic due to the long chain of polar words, which can lead to the biased polarity value of sentence. Moreover, the positive-negative word pairs need to be in the single sentence (i.e not separated by the conjunction) and the weight of the emoticons and #tag are nullified which are considered here as a fundamental feature for the sarcasm detection. The experiment results significantly dropped when longer and complex sentences are used. So, the necessity of more robust methods is required to detect the sarcasm. Majority of research is done using either Machine Learning and Lexical based approach using either dictionary or corpus based and the measure is statistical formula. However, further exploration in area of semantic information embedding should done in order to detect correct polarity value of word in particular context. These context dependent values eventually help to improve theresults of sarcasm detection.

In our future work we are focusing on the Neural network and application of the Deep learning model on the textual data that we have collected and measure performance indexes of the these models on dataset and prepare the comparison on the same parameter including the length of sentences and presence of the other features.

References

1. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.
2. Joshi, Aditya, A. R. Balamurali, and Pushpak Bhattacharyya. "A fall-back strategy for sentiment analysis in Hindi: a case study." Proceedings of the 8th ICON (2010).
3. Jiang, Long, et al. "Target-dependent twitter sentiment classification." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.
4. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.
5. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
6. Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005.
7. Read, Jonathon. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification." Proceedings of the ACL student research workshop. Association for Computational Linguistics, 2005.
8. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.
9. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.
10. Narr, Sascha, Michael Hulfenhaus, and Sahin Albayrak. "Language-independent twitter sentiment analysis." Knowledge Discovery and Machine Learning (KDML), LWA (2012): 12- 14.
11. Maynard, Diana, Kalina Bontcheva, and Dominic Rout. "Challenges in developing opinion mining tools for social media." Proceedings of the@ NLP can u tag# usergeneratedcontent (2012): 15-22.
12. Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.
13. Nigam, Kamal, John Lafferty, and Andrew McCallum. "Using maximum entropy for text classification." IJCAI-99 workshop on machine learning for information filtering. Vol. 1. 1999.
14. Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

15. Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
16. González-Ibánez, Roberto, Smaranda Muresan, and Nina Wacholder. "Identifying sarcasm in Twitter: a closer look." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011.
17. Liebrecht, C. C., F. A. Kunneman, and A. P. J. van den Bosch. "The perfect solution for detecting sarcasm in tweets# not." (2013).
18. Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Semi- supervised recognition of sarcastic sentences in twitter and amazon." Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2010.
19. Riloff, Ellen, et al. "Sarcasm as Contrast between a Positive Sentiment and Negative Situation." EMNLP. 2013.
20. Rajadesingan, Ashwin, Reza Zafarani, and Huan Liu. "Sarcasm detection on Twitter: A behavioral modeling approach." Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015.
21. Bamman, David, and Noah A. Smith. "Contextualized Sarcasm Detection on Twitter." Ninth International AAAI Conference on Web and Social Media. 2015
22. Smith, Phillip, et al. "Sentiment Analysis: Beyond Polarity." (2011).
23. Lunando, Edwin, and Ayu Purwarianti. "Indonesian social media sentiment analysis with sarcasm detection." Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on. IEEE, 2013.
24. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014): 1093-1113.
25. Tsytsarau, Mikalai, and Themis Palpanas. "Survey on mining subjective data on the web." Data Mining and Knowledge Discovery 24.3 (2012): 478-514.Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18.5 (2011): 544-551.
26. Aggarwal, Charu C., and ChengXiang Zhai. Mining text data. Springer Science & Business Media, 2012.
27. Piyush Arora, a thesis on “Sentiment Analysis For Hindi Language”.
28. World Language and Culture: Top 30 Languages by Number of Native Speakers (2005). http://www.vistawide.com/languages/top_30_languages.ht m
29. Pushpak Bhattacharyya , “Sarcasm Detection: A Computational and Cognitive Study”
30. Survey : “Opinion Mining and Sentiment Analysis”
31. Dave, A. D., & Desai, N. P. (2016, March). A comprehensive study of classification techniques for sarcasm detection on textual data. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 1985-1991).
32. Desai N, Dave AD. Sarcasm detection in hindi sentences using support vector machine. International Journal. 2016 Jul;4(7):8-15.
33. Bharti SK, Babu KS, Jena SK. Harnessing online news for sarcasm detection in hindi tweets. InInternational Conference on Pattern Recognition and Machine Intelligence 2017 Dec 5 (pp. 679-686). Springer, Cham.
34. Ebrahimi M, Yazdavar AH, Sheth A. Challenges of sentiment analysis for dynamic events. IEEE Intelligent Systems. 2017 Sep;32(5):70-5.