How Do Cultural Differences Impact the Quality of Sarcasm Annotation?: A Case Study of Indian Annotators and American Text

Aditya Joshi[1,2,3] **Pushpak Bhattacharyya**[1] **Mark Carman**[2]
Jaya Saraswati[1] **Rajita Shukla**[1]
[1]IIT Bombay, India
[2]Monash University, Australia
[3]IITB-Monash Research Academy, India
{adityaj, pb}@cse.iitb.ac.in, mark.carman@monash.edu

Abstract

Sarcasm annotation extends beyond linguistic expertise, and often involves cultural context. This paper presents our first-of-its-kind study that deals with impact of cultural differences on the quality of sarcasm annotation. For this study, we consider the case of American text and Indian annotators. For two sarcasm-labeled datasets of American tweets and discussion forum posts that have been annotated by American annotators, we obtain annotations from Indian annotators. Our Indian annotators agree with each other more than their American counterparts, and face difficulties in case of unfamiliar situations and named entities. However, these difficulties in sarcasm annotation result in **statistically insignificant degradation** in sarcasm classification. We also show that these disagreements between annotators can be predicted using textual properties. Although the current study is limited to two annotators and one culture pair, our paper opens up a novel direction in evaluation of the quality of sarcasm annotation, and the impact of this quality on sarcasm classification. This study forms a stepping stone towards systematic evaluation of quality of these datasets annotated by non-native annotators, and can be extended to other culture combinations.

1 Introduction

Sarcasm is a linguistic expression where literal sentiment of a text is different from the implied sentiment, with the intention of ridicule (Schwoebel et al., 2000). Several data-driven approaches have been reported for computational detection of sarcasm (Tsur et al., 2010; Davidov et al., 2010; Joshi et al., 2015). As is typical of supervised approaches, they rely on datasets labeled with sarcasm. We refer to the process of creating such sarcasm-labeled datasets as sarcasm annotation.

Linguistic studies concerning cross-cultural dependencies of sarcasm have been reported (Boers, 2003; Thomas, 1983; Tannen, 1984; Rockwell and Theriot, 2001; Bouton, 1988; Haiman, 1998; Dress et al., 2008). However, these studies do not look at the notion of cross-cultural sarcasm *annotation* of text. This paper reports the first set of findings from our ongoing line of research: evaluation of quality of sarcasm annotation when obtained from annotators of non-native cultures.

We consider the case of annotators of Indian origin annotating datasets (consisting of discussion forums/tweets from the US) that were earlier annotated by American annotators. It may be argued that since crowd-sourcing is prevalent now, a large pool of annotators makes up for cultural differences among few annotators. However, a fundamental study like ours that performs a micro-analysis of culture combinations is likely to be useful for a variety of reasons such as judging the quality of new datasets, or deciding among annotators. Balancing the linguistic and computational perspectives, we present our findings in two ways: (a) degradation in quality of sarcasm annotation by non-native annotators, and (b) impact of this quality on sarcasm classification.

The motivation behind our study is described in Section 2, while our annotation experiments are in Section 3. We present our analysis in terms of four questions: (a) Are there peculiar difficulties that non-native annotators face during sarcasm annotation? (Section 4.1), (b) How do these difficulties impact the quality of sarcasm annotation? (Section 4.2), (c) How do cultural differences affect sarcasm classification that uses such annotation? (Section 4.3), and (c) Can these difficulties be predicted using features of text? (Section 4.4). All labeled datasets are available on request for future work. For every textual unit, they contain multiple annotations, by native (as given in past works), and non-native annotators.

2 Why is such an evaluation of quality important?

To build NLP systems, creation of annotated corpora is common. When annotators are hired, factors such as language competence are considered. However, while tasks like sense annotation or part-of-speech labeling require linguistic expertise, sarcasm annotation extends beyond linguistic expertise, and often involves cultural context. Tannen (1984) describe how a guest thanking

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 95–99,
Berlin, Germany, August 11, 2016.

the host for a meal may be perceived as polite in some cultures, but sarcastic in some others.

Due to popularity of crowd-sourcing, cultural background of annotators may not be known at all. Keeping these constraints in mind, a study of non-native annotation, and its effect on the corresponding NLP task assumes importance. Our work is the first-of-its-kind study related to sarcasm annotation. Similar studies have been reported for related tasks. Hupont et al. (2014) deal with result of cultural differences on annotation of images with emotions. Das and Bandyopadhyay (2011) describe multi-cultural observations during creation of an emotion lexicon. For example, they state that the word 'blue' may be correlated to sadness in some cultures but to evil in others. Similar studies to understand annotator biases have been performed for subjectivity annotation (Wiebe et al., 1999) and machine translation (Cohn and Specia, 2013). Wiebe et al. (1999) show how some annotators may have individual biases towards a certain subjective label, and devise a method to obtain bias-corrected tags. Cohn and Specia (2013) consider annotator biases for the task of assigning quality scores to machine translation output.

3 Our Annotation Experiments

In this section, we describe our annotation experiments in terms of datasets, annotators and experiment details.

3.1 Datasets

We use two sarcasm-labeled datasets that have been reported in past work. The first dataset is **Tweet-A**. This dataset, introduced by Riloff et al. (2013), consists of 2278 manually labeled tweets, out of which 506 are sarcastic. We call these annotations American1. An example of a sarcastic tweet in this dataset is '*Back to the oral surgeon #yay*'. The second dataset is **Discussion-A**: This dataset, introduced by Walker et al. (2012), consists of 5854 discussion forum posts, out of which 742 are sarcastic. This dataset was created using Amazon Mechanical Turk. IP addresses of Turk workers were limited to USA during the experiment[1]. We call these annotations American2. An example post here is: '*A master baiter like you should present your thesis to be taken seriously. You haven't and you aren't.*'.

3.2 Our Annotators

Our annotators are two female professional linguists of Indian origin with more than 8K hours of experience in annotating English documents for tasks such as sentiment analysis, word sense disambiguation, etc. [2]. They are both 50+ years old and follow only international news that would expose them to American culture. We refer to these annotators as Indian1 and Indian2. The choice of 'Indian' annotators was made bearing in mind the difference between American and Indian cultures. Our two-annotator configuration is reasonable due to explanation provided in Section 1. Also, it is similar to Riloff et al. (2013) where three annotators create a sarcasm-labeled dataset.

3.3 Experiments

The annotation experiment is conducted as follows. Our annotators read a unit of text, and determine whether it is sarcastic or not. The experiment is conducted in sessions of 50 textual units, and the annotators can pause anywhere through a session. This results in datasets where each textual unit has three annotations as follows: (A) Tweet-A annotated by American1, Indian1, Indian2, (B) Discussion-A annotated by American2, Indian1, Indian2. The American annotations are from past work. (A) and (B) differ in domain (tweets v/s discussion forum posts). These annotated datasets are available on request.

4 Analysis

We now analyze these datasets from three perspectives: (a) difficulties during creation and impact on quality, (b) degradation in annotation quality, (c) impact of quality degradation on sarcasm classification, and (c) prediction of disagreement.

4.1 What difficulties do our Indian annotators face?

Table 1 shows examples where our Indian annotators face difficulty in annotation. We describe experiences from the experiments in two parts:

1. **Situations in which they were unsure of the label**: These include sources of confusion for our annotators, but may or may not have led to incorrect labels.

   Data bias: There are more non-sarcastic texts in the dataset than sarcastic ones. Despite that, the annotators experienced suspicion about every sentence that they had to mark as sarcastic or non-sarcastic. This resulted in confusion as in the case of example *1* in Table 1.
   Unfamiliar words: The annotators consult a dictionary for jargon like 'abiogenesis' or 'happenstance'. For urban slang, they look up the urban dictionary website[3]. Hashtags and emoticons were key clues that the annotators used to detect sarcasm. For example, '*No my roommate play out of tune Zeppelin songs right outside my door isnt annoying. Not at all #sigh*'. They also stated that they could understand the meaning after few occurrences. They had to verify the annotation that they had assigned in the previous instances. Thus, it is helpful *if annotation tools*

[1] We acknowledge the possibility that some of these annotators where not physically located within USA, despite IP, due to VPN or similar infrastructure

[2] This description highlights that they have strong linguistic expertise.

[3] http://www.urbandictionary.com/

	Example	Remarks
	Situations in which they were unsure of the label	
1	I have the worlds best neighbors!	The annotators were not sure if this was intended to be sarcastic. Additional context would have been helpful.
	Situations in which their label did not match that by American annotators	
2	@twitter_handle West Ham with Carlton Cole and Carroll up front. Going to be some free flowing football this season then	Annotators were not familiar with these players. Hence, they were unable to determine the underlying sentiment.
3	And, I'm sure that Terri Schiavo was fully aware of all that Bush and radical right-wing religionists did for her and appreciates what they did.	Indian annotators did not know about Terri Schiavo, and had to look up her story on the internet.
4	Love going to work and being sent home after two hours	The Indian annotators were unaware of the context of the long commute and the possibility that 'being sent home' meant being fired from job. Hence, they could not detect the sarcasm.
5	@twitter_handle Suns out and I'm working,#yay	The annotators were not sure if a sunny day is pleasant - considering temperatures in India.
6	'So how are those gun free zones working out for you?'	With inadequate knowledge about gun free zones, the annotators were doubtful about sarcasm in the target sentence.

Table 1: Examples of sentences that the Indian annotators found difficult to annotate; 'twitter_handle' are twitter handles suppressed for anonymity

Annotator Pair	κ	Agreement (%)
Avg. American1	0.81	-
Indian1 & Indian2	**0.686**	85.82
Indian1 & American1	0.524	80.05
Indian2 & American1	0.508	79.98

Table 2: Inter-annotator agreement statistics for Tweet-A; Avg. American1 is as reported in the original paper

allow correction of a previously annotated text, since annotators may understand such words during the course of annotation.

2. **Situations in which their labels did not match their American counterparts**:

   Unknown context about named entities Consider examples *2* and *3* in Table 1. In case of named entities in domains such as sports or politics, annotators were unfamiliar with popular figures and their associated sentiment. **Unknown context about situations**: Example *4* is a case of Indian annotators marking a text as non-sarcastic, while their American counterparts marked it as sarcastic. **Unclear understanding of socio-political situations**: The tweet in example *5* was labeled as non-sarcastic by Indian annotators. Similarly, example *6* appears to be a non-sarcastic question. However, based on their perception about gun shooting incidents in USA, they were unsure if this statement was indeed non-sarcastic.

4.2 How do cross-cultural difficulties affect quality of annotation?

We now compare quality of non-native annotation using inter-annotator agreement metrics. Table 2 shows statistics for Tweet-A dataset. Kappa coefficient as reported in the original paper is 0.81. The corresponding value between Indian1 and Indian2 is 0.686. The values for discussion forum dataset Discussion-A are shown in Table 4. For Discussion-A, Kappa coefficient between the two Indian annotators is 0.700, while that between Indian1/2 and American annotators is 0.569 and 0.288 respectively. Average values for American annotators are not available in the original paper, and hence not mentioned. This shows that inter-annotator agreement between our annotators is higher than their individual agreement with the American annotators. Kappa values are lower in case of tweets than discussion forum posts. Agreement (%) indicates the percentage overlap between a pair of labels. This agreement is high between Indian annotators in case of Tweet-A (85.82%), and Discussion-A (92.58%), and comparable with American annotators.

Annotator Pair	κ	Agreement (%)
Indian1 & Indian2	**0.700**	92.58
Indian1 & American2	0.569	89.81
Indian2 & American2	0.288	83.33

Table 4: Inter-annotator agreement statistics for Discussion-A

Table 5 shows the percentage agreement separately for the two classes, with American labels as reference labels. In case of Tweet-A, our annotators agree more with American annotators on sarcastic than non-sarcastic tweets. This means that in case of short text such as tweets, it is the non-sarcastic tweets that cause disagreement. This highlights the fuzziness of sarcastic expressions. On the contrary, in case of long text such as discussion forum posts, sarcastic tweets cause disagreement for our annotators because sarcasm may be in a short portion of a long discussion forum post.

Training Label Source	Test Label Source	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)	AUC
			Tweet-A			
American	American	80.5	71.5	69.2	70.27	0.816
Indian	American	74.14	65.78	68.61	65.28	0.771
			Discussion-A			
American	American	83.9	61.5	59.05	59.97	0.734
Indian	American	79.42	58.28	56.77	56.36	0.669

Table 3: Impact of non-native annotation on sarcasm classification; Values for Indian-American are averaged over Indian annotators

Annotator Pair	Sarcastic	Non-sarc
	Tweet-A	
Indian1 & American1	**84.78**	77.71
Indian2 & American1	79.24	**80.24**
	Discussion-A	
Indian1 & American2	67.24	**93**
Indian2 & American2	40.91	**89.5**

Table 5: Class-wise agreement (%) for pairs of annotators, for both datasets

4.3 How do these difficulties affect sarcasm classification?

We now evaluate if difficulties in sarcasm annotation have an impact on sarcasm classification. To do so, we use LibSVM by Chang and Lin (2011) with a linear kernel to train a sarcasm classifier that predicts a given text as sarcastic or not. We use unigrams as features, and report five-fold cross-validation performance. Table 3 shows performance values for Discussion-A and Tweet-A, specifically, Accuracy, Precision, Recall, F-score and Area Under Curve (AUC). These values are averaged over both Indian annotators, for the respective configuration of training labels[4]. For Tweet-A, using the dataset annotated by American annotators as training labels, leads to an AUC of 0.816. The corresponding value when annotation by Indian annotators is used, is 0.771. Similar trends are observed in case of other metrics, and also for Discussion-A. However, *degradations for both Tweet-A and Discussion-A are not statistically significant for the 95% confidence interval*. Thus, although our Indian annotators face difficulties during annotation resulting in partial agreement in labels, it seems that annotations from these annotators did not lead to significant degradation to what the sarcasm annotation will eventually be used for, *i.e.*, sarcasm classification. The two-tailed p-values for Tweet-A and Discussion-A are 0.221 and 0.480 respectively.

[4]This means that the experiment in case of Indian annotators as training labels consisted of two runs, one for each annotator.

4.4 Can disagreements be predicted?

We now explore if we can predict, solely using properties of text, whether our Indian annotators will disagree with their American counterparts. This goal is helpful so as to choose between annotators for a given piece of text. For example, if it can be known beforehand (as we do in our case) that a text is likely to result in a disagreement between native and non-native annotators, its annotation can be obtained from native annotator alone. With this goal, we train a SVM-based classifier that predicts (dis)agreement. In the training dataset, the agreement label is assigned using our datasets with multiple annotations. We use three sets of features: (a) POS, (b) Named entities, (c) Unigrams (a & b are obtained from NLTK (Bird, 2006)). Table 6 shows performance for 3-fold cross-validation, averaged over the two annotators as in the previous case. We obtain an AUC of 0.56 for Tweet-A, and 0.59 for Discussion-A. The high accuracy and AUC values show that words and lexical features (such as named entities and part-of-speech tags) can effectively predict disagreements.

Dataset	Accuracy (%)	AUC
Tweet-A	67.10	**0.56**
Discussion-A	75.71	**0.59**

Table 6: Predicting annotator agreement using textual features; Values are averaged over Indian annotators

5 Conclusion & Future Work

Concerns about annotation quality may be raised if nature of the task is dependent on cultural background of annotators. In this paper, we presented a first-of-its-kind annotation study that evaluates quality of sarcasm annotation due to cultural differences. We used two datasets annotated by American annotators: one consisting of tweets, and another consisting of discussion forum posts. We obtained another set of sarcasm labels from two annotators of Indian origin, similar to past work where three annotators annotate a dataset with sarcasm labels. We discussed our findings in three

steps. The key insights from each of these steps are as follows: (1) Our Indian annotators agree with each other more than they agree with their American counterparts. Also, in case of short text (tweets), the agreement is higher in sarcastic text while for long text (discussion forum posts), it is higher in non-sarcastic text. Our annotators face difficulties due to unfamiliar situations, named entities, etc. (2) Our sarcasm classifiers trained on labels by Indian annotators show a statistically insignificant (as desired) degradation as compared to trained on labels by American annotators, for Tweet-A (AUC: 0.816 v/s 0.771), and for Discussion-A (AUC: 0.734 v/s 0.669). (3) Finally, using textual features, the disagreement/difficulty in annotation can be predicted, with an AUC of 0.56.

Sarcasm detection is an active research area, and sarcasm-labeled datasets are being introduced. Our study forms a stepping stone towards systematic evaluation of quality of these datasets annotated by non-native annotators, and can be extended to other culture combinations.

References

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Computational Linguistics 2006: Systems Demonstrations*, pages 69–72.

Frank Boers. 2003. Applied linguistics perspectives on cross-cultural variation in conceptual metaphor. *Metaphor and Symbol*, 18(4):231–238.

Lawrence F Bouton. 1988. A cross-cultural study of ability to interpret implicatures in english. *World Englishes*, 7(2):183–196.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Computational Linguistics 2013*, pages 32–42.

Amitava Das and Sivaji Bandyopadhyay. 2011. Dr sentiment knows everything! In *Proceedings of the Annual Meeting of the Association for Computational Linguistics/Human Language Technologies 2011: Systems Demonstrations*, pages 50–55.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning 2010*, pages 107–116.

Megan L Dress, Roger J Kreuz, Kristen E Link, and Gina M Caucci. 2008. Regional variation in the use of sarcasm. *Journal of Language and Social Psychology*, 27(1):71–85.

John Haiman. 1998. Talk is cheap: Sarcasm, alienation, and the evolution of language. *Oxford University Press*.

Isabelle Hupont, Pierre Lebreton, Toni Maki, Evangelos Skodras, and Matthias Hirth. 2014. Is affective crowdsourcing reliable? In *IEEE Fifth International Conference on Communications and Electronics (ICCE) 2014*, pages 516–521.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2015*, page 757.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2013*, pages 704–714.

Patricia Rockwell and Evelyn M Theriot. 2001. Culture, gender, and gender mix in encoders of sarcasm: A self-assessment analysis. *Communication Research Reports*, 18(1):44–52.

John Schwoebel, Shelly Dews, Ellen Winner, and Kavitha Srinivas. 2000. Obligatory processing of the literal meaning of ironic utterances: Further evidence. *Metaphor and Symbol*, 15(1-2):47–61.

Deborah Tannen. 1984. The pragmatics of cross-cultural communication. *Applied Linguistics*, 5(3):189–195.

Jenny Thomas. 1983. Cross-cultural pragmatic failure.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *International AAAI Conference on Web and Social Media 2010*.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Language Resources and Evaluation Conference 2014*, pages 812–817.

Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Computational Linguistics 1999*, pages 246–253.