Predicting Readers' Sarcasm Understandability by Modeling Gaze Behavior

Abhijit Mishra, Diptesh Kanojia, Pushpak Bhattacharyya
Center for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India,
{abhijitmishra, diptesh, pb}@cse.iitb.ac.in

Abstract

Sarcasm understandability or the ability to understand textual sarcasm depends upon readers' language proficiency, social knowledge, mental state and attentiveness. We introduce a novel method to predict the sarcasm understandability of a reader. Presence of *incongruity* in textual sarcasm often elicits distinctive eye-movement behavior by human readers. By recording and analyzing the eye-gaze data, we show that eye-movement patterns vary when sarcasm is understood *vis-à-vis* when it is not. Motivated by our observations, we propose a system for sarcasm understandability prediction using supervised machine learning. Our system relies on readers' eye-movement parameters and a few textual features, thence, is able to predict sarcasm understandability with an F-score of 93%, which demonstrates its efficacy.

The availability of inexpensive embedded-eye-trackers on mobile devices creates avenues for applying such research which benefits web-content creators, review writers and social media analysts alike.

1 Introduction

Sarcasm is an intensified and complex way of expressing a negative remark that involves *mocking, contemptuous, or ironic language*[1]. Understanding it demands carefully orchestrated sequences of complicated cognitive activities in the brain (Shamay, Tomer, and Aharon 2005). This may depend on readers' language proficiency, social knowledge, mental state and attentiveness while reading textual sarcasm. Failure in understanding sarcasm can be attributed to lack of any of these factors.

Can machines predict whether a reader has understood the intended meaning of a sarcastic text? We refer to this problem as *Sarcasm Understandability Prediction*. The importance of this problem can be felt in multiple scenarios such as: (a) *Online review construction and analysis*, where knowing sarcasm understandability of the target audience can help prepare and organize reviews more effectively, (b) *Language learning*, say, monitoring the progress of a second language learner where sarcasm understandability can be a factor in determining the level of proficiency, and, (c) *Attentiveness testing*, where readers, especially children, learning from online courses can be instructed to be more attentive if they show impatience while reading.

[1]Source: The Free Dictionary

We introduce a novel way of predicting the *sarcasm understandability* of a reader. Our proposed system takes **readers' eye-gaze parameters as input along with textual features to determine whether the reader has understood the underlying sarcasm or not**. This way of addressing the problem is plausible due to two reasons: (a) Cognitive processes in the brain are related to eye-movement activities (Parasuraman and Rizzo 2006). Hence, considering readers' eye-movement patterns for detection of *sarcasm understandability* offers a more natural setting that does not interrupt the reader. This is unlike an explicit mode of evaluation ,say, by judging through questionnaires. (b) Availability of inexpensive embedded eye-trackers on hand-held devices has come close to reality now. For instance, *Cogisen*[2] has a patent (ID: EP2833308-A1) on *eye-tracking using inexpensive mobile web-cameras*. Thus, we can soon expect to gather the eye-movement data of a large number of online readers. This builds a strong use-case for our application.

Terminology: In the subsequent sections, we use the terms *sarcasm_hit* and *sarcasm_miss*, for the conditions of sarcasm being understood and sarcasm not being understood by the reader.

2 Related Work

Sarcasm processing and understanding has been studied for quite some time. A few pioneering works in this area include that of Jorgensen, Miller, and Sperber (1984), who believe that sarcasm arises when a figurative meaning is used opposite to the literal meaning of the utterance. According to Clark and Gerrig (1984), sarcasm processing involves canceling the indirectly negated message and replacing it with the implicated one. Giora (1995), on the other hand, define sarcasm as a mode of indirect negation that requires processing of both negated and implicated messages. Ivanko and Pexman (2003) study the inherent complexity of sarcasm and its effect on sarcasm processing time. From the computational linguistic perspective, several automatic sarcasm detection systems have been proposed and implemented using rule based and statistical techniques considering *(a) Unigrams and Pragmatic features* (Carvalho et al. 2009; González-Ibánez, Muresan, and Wacholder 2011; Barbieri, Saggion, and Ronzano 2014; Joshi, Sharma, and Bhattacharyya 2015) *(b) Stylistic patterns* (Davidov, Tsur, and Rappoport 2010; Riloff et al. 2013) *(c) Tweet hashtag inter-*

[2]http://www.sencogi.com

pretations (Liebrecht, Kunneman, and van den Bosch 2013; Maynard and Greenwood 2014). Under different settings, the best accuracies (in terms of F-score) of these systems vary between 60% to 90%.

With the advent of sophisticated eye-trackers and electro/magneto-encephalographic (EEG/MEG) devices, it has been possible to delve deep into the cognitive underpinnings of sarcasm understanding. Shamay, Tomer, and Aharon (2005) perform a neuroanatomical analysis of sarcasm understanding by observing the performance of participants with focal lesions on tasks that required understanding of sarcasm and social cognition. Camblin, Gordon, and Swaab (2007) show that in multi-sentence passages, discourse congruence has robust effects on eye movements. This also implies that disrupted processing occurs for discourse incongruent words, even though they are perfectly congruous at the sentence level. Filik (2014), using a series of eye-tracking and EEG experiments, find that, for unfamiliar ironies, the literal interpretation would be computed first, and a mismatch with context would lead to a re-interpretation of the statement as being ironic.

Reading researchers have applied eye-tracking for behavioral studies as surveyed and summarized by Rayner (1998). But, from a computational perspective, there are less number of works that quantify or predict various levels of difficulties associated with reading comprehension and text understanding ability. Martınez-Gómez and Aizawa (2013) quantify reading difficulty using readers' eye-gaze patterns with the help of Bayesian Learning. Mishra, Bhattacharyya, and Carl (2013) propose a framework to predict difficulty in text translation using translator's eye-gaze patterns with the help of supervised machine learning approach. Similarly, Joshi et al. (2014) introduce a system for measuring the difficulties perceived by humans in understanding the sentiment expressed in texts.

Our method of analyzing eye-movement data for *sarcasm understandability* is the first of its kind and is inspired by these recent advancements.

3 Sarcasm, Cognition and Eye-movement

Sarcasm often emanates from context incongruity (Campbell and Katz 2012), which, possibly, surprises the reader and enforces a re-analysis of the text. The time taken to understand sarcasm (referred to as the *Sarcasm Processing Time*) depends on the *degree of context incongruity* between the statement and the context (Ivanko and Pexman 2003). In the absence of any information regarding the nature of the forthcoming text, intuitively, human brain would start processing the text in a sequential manner, with the aim of comprehending the literal meaning. When incongruity is perceived, the brain may initiate a re-analysis to reason out such disparity (Kutas and Hillyard 1980) . *As information during reading is passed to brain through eyes, incongruity may affect the way eye-gaze moves through the text. Hence, distinctive eye-movement patterns may be observed in the case of successful processing of sarcasm, in contrast to an unsuccessful attempt.*

This hypothesis forms the crux of our analysis and we aim to prove/disprove this by creating and analyzing an eye-movement database for sarcasm reading. Our database can be freely downloaded[3] for academic purposes.

4 Creation of Eye-movement Database

The experimental setup for the collection of eye-movement database is described below.

4.1 Document Description

We prepared a database of 1,000 short-text, comprising 10-40 words and one or two sentences. Out of these texts, 350 are sarcastic and were collected as follows: (a) 103 sentences were manually extracted from two popular sarcastic quote websites[4], (b) 76 sarcastic short movie reviews were manually extracted from the *Amazon Movie Corpus* (Pang and Lee) and (c) 171 tweets were downloaded using the hashtag *#sarcasm* from Twitter. The 650 non-sarcastic texts were either downloaded from Twitter or extracted from the Amazon Movie Review corpus. The sentences do not contain *highly* topic/culture/country specific phrases. The tweets were normalized to avoid difficulty in interpreting social media lingo. All the sentences in our dataset carry either positive or negative opinion about specific "aspects". For example, the sentence *"The movie is extremely well cast"* has positive sentiment about the aspect "cast".

Two expert linguists validated with 100% agreement that the 350 sentences extracted from various sources are indeed sarcastic and express *negative* sentiment towards the main aspect. This forms the ground truth for our experiment.

4.2 Participant Description

We chose seven graduate students with science and engineering background in the age group of 22-27 years with English as the primary language of academic instruction. Our participants are non-native speakers of English. We confirm that their native languages are read from left-to-right. This eliminates the possibility of our experiment getting affected by reading direction.

To ensure that they possess excellent level of proficiency in English, our participants are selected based on their ToEFL-iBT scores of 100 or above. They are given a set of instructions beforehand, that mention the nature of the task, annotation input method, and necessity of head movement minimization during the experiment. Participants were financially rewarded for their effort.

Though our analysis is based on the current observations involving non-native speakers (with acceptable English proficiency), our predictive framework is *data driven* and does not rely heavily on any assumption about the nature of the eye-movement patterns. Additionally, we have used *linguistic* and *readability* related features to ensure that, the combination of eye-gaze and linguistic/readability parameters will be able to discriminate between *sarcasm_hit* and *sarcasm_miss* cases. So, our approach is expected to work for a general population of both native and non-native speakers.

[3] http://www.cfilt.iitb.ac.in/cognitive-nlp/

[4] http://www.sarcasmsociety.com, http://www.themarysue.com/funny-amazon-reviews

	All		Quote		Twitter		Movie	
	Acc.	IAA	Acc.	IAA	Acc.	IAA	Acc.	IAA
P1	79.71	0.71	81.73	0.72	76.74	0.69	81.58	0.87
P2	89.14	0.77	83.65	0.74	90.12	0.80	92.11	0.76
P3	87.14	0.75	86.54	0.75	88.37	0.76	82.89	0.71
P4	88	0.76	87.5	0.77	88.37	0.77	85.53	0.74
P5	72.57	0.65	78.85	0.70	67.44	0.62	73.68	0.64
P6	90.29	0.77	85.58	0.75	95.35	0.81	82.89	0.72
P7	85.71	0.75	81.73	0.73	87.79	0.77	84.21	0.73

Table 1: Annotation results for seven participants. *Acc.* $\rightarrow$ Percentage of sarcastic sentences correctly identified. *IAA* $\rightarrow$ Average Cohen's Kappa Inter Annotator Agreement between a participant and others

4.3 Task Description

The task assigned to our participants is to read one sentence at a time and annotate with binary sentiment polarity labels (*i.e.,* positive/negative). Note that we do not instruct our participants to explicitly annotate whether a sentence is sarcastic or not. It has been shown by Gibbs (1986) that processing incongruity becomes relatively easier if sarcasm is expected beforehand. But, in practice, it seldom occurs that a reader has prior knowledge about the nature of the forthcoming text. Our setup ensures "ecological validity" in two ways. (1) Readers are not given any clue that they have to treat sarcasm with special attention. This is done by setting the task to polarity annotation (instead of sarcasm detection). (2) Sarcastic sentences are mixed with non sarcastic text, which does not give prior knowledge about whether the forthcoming text will be sarcastic or not.

The eye-tracking experiment is conducted by following the standard norms in eye-movement research (Holmqvist et al. 2011). At a time, one sentence is displayed to the reader along with the "aspect" with respect to which the annotation has to be provided. While reading, an SR-Research Eyelink-1000 eye-tracker (monocular remote mode, sampling rate 500Hz) records several eye-movement parameters like gaze-fixations (*i.e.*, a long stay of gaze), saccade (*i.e.*, quick jumping of gaze between two positions of rest) and pupil size. The whole experiment is divided into 20 sessions, each having 50 sentences to be read and annotated. This is to prevent fatigue over a period of time. However, there was no time limit on individual sentence annotations.

For our analysis, we consider only the 350 sarcastic sentences in our data-set, since our system requires the sentences to be sarcastic. We acknowledge that the analysis of understandability of other linguistically complex forms of text could be carried out using the rest of the 650 sentences. This is, however, beyond the scope of this work.

Even though we took all the necessary measures to control the experiment, the individual annotation accuracy for our participants is still far from being 100%. (However, the agreement between consensus and gold annotations is 98%, ensuring the sanctity of the gold data). This shows the inherent difficulty of sarcasm understanding. Table 1 shows the annotation accuracy of the seven participants along with the average *Inter Annotator Agreement* of each participant (P1-P7) with others, separately shown for the whole dataset and individual domains. The incorrectness in annotation can be attributed to: (a) Lack of patience/attention while reading, (b) Issues related to text comprehension and understandability, and (c) Confusion/indecisiveness caused due to lack of context.

Our objective is to find out if a reader has understood sarcasm or not. How can polarity annotation help here? We assume that if the reader does not understand sarcasm, the text will be annotated with an incorrect polarity label. Our whole analysis relies on this **key-assumption**.

5 Analysis of Eye-movement Data

To show the dependencies between reader's eye movement patterns and sarcasm understandability, we perform a two-fold analysis of the eye-movement data intending to observe: (a) Variation in *eye-gaze attributes* and (b) Variation in *scan-paths*.

5.1 Variation in Eye-gaze Attributes

Eye-movement patterns are characterized by two basic attributes: (1) Fixations, corresponding to a longer stay of gaze on a visual object (like characters, words *etc.* in text) (2) Saccades, corresponding to the transition of eyes between two fixations. Moreover, a saccade is called a *Regressive Saccade* or simply, *Regression* if it represents a phenomenon of going back to a pre-visited segment. A portion of a text is said to be *skipped* if it does not have any fixation.

We perform a series of *t-tests*[5] to check if there is a statistically significant difference between the distributions of various gaze attributes across all participants for the conditions of *sarcasm_miss* and *sarcasm_hit*. We consider four basic gaze attributes, *viz.*, (1) Average fixation duration per word (2) Average count of fixations per word (3) Total regression counts per word and (4) Percentage of words skipped. These attributes are taken from *reading literature* for behavioral studies (Rayner 1998). The null hypothesis for each gaze attribute is: There should not be any significant difference between the *mean* (μ) of the attribute for both *sarcasm_miss* and *sarcasm_hit*. The threshold α for accepting or rejecting the null hypothesis is set to *0.05*. The test results are shown in Table 2. Except for *average count of fixations*, all other attributes exhibit significant differences, with *fixation duration* having the highest difference.

It is evident from the tests that the average time spent (in terms of fixation duration) is generally higher for *sarcasm_miss* than for *sarcasm_hit*. For both *fixation duration* and *regressive saccade count*, we consistently observe a higher variance for *sarcasm_miss*. This observation can be intuitively felt by considering the following scenarios: *(a)* When sarcasm is not understood because of lack of attentiveness, the reader may spend less time/focus on different segments of the text. Moreover, more words may be skipped and the number of regressive saccades will be less. *(b)* When sarcasm is not understood because of lack of *linguistic expertise, conceptual knowledge or cognitive ability*, the reader may spend fixate more on different segments. The

[5] two-tailed assuming unequal variance

Gaze-Attribute	$\mu_{sarcasm\_miss}$	$\sigma_{sarcasm\_miss}$	$\mu_{sarcasm\_hit}$	$\sigma_{sarcasm\_hit}$	t	p	Remark
Fixation Duration (in ms)	357	229	307	176	3.8	0.00017	Significant
Average Count of Fixations	0.79	0.2	0.77	0.18	1.89	0.05	Not significant
Count of Regressive Saccades	6	5	3.2	2.79	2.27	0.01	Significant
Word Skip Percentage	28.6	14.6	30.1	15.1	-2.06	0.03	Significant

Table 2: T-test statistics for different eye-movement attributes for the conditions of (i) *sarcasm_miss* and (ii) *sarcasm_hit*

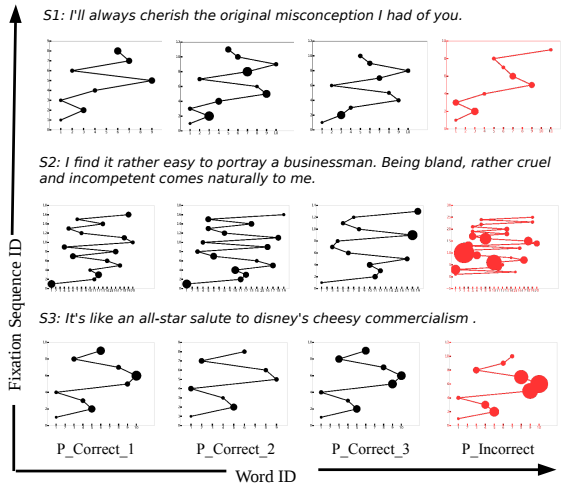

Figure 1: Scanpaths of four different sets of participants for three sarcastic sentences S1, S2 and S3. The circles represent fixations, edges represent saccades and areas of the circle represent fixation duration. X- and Y- axes respectively represent the positions of words in the sentence and temporal sequence in which the fixations occur. Scanpaths of participants who have not identified sarcasm correctly ($P_{Incorrect}$) are shown in red.

skipping rate will be low and the reader's eye may re-visit various segments from time to time, perhaps, to gather more clues for interpreting the text. Hence, more regressive saccades may be observed.

Note that our observations of variations in different gaze attributes may not be confined to the presence of sarcasm. For example, a *garden-path* sentence like "The horse raced past the barn fell" may enforce the brain to perform a syntactic reanalysis of the text. This may affect gaze-fixations and regressions (Malsburg and Vasishth 2011) the same way as sarcasm does. But, we rule out such cases as our objective is to predict understandability for sarcastic texts.

We extend our analysis to gain more insight into the eye-movement behavior, and its relation with sarcasm understandability by considering *scanpaths*.

5.2 Variation in Scanpaths

Scanpaths are line-graphs that contain fixations as nodes and saccades as edges; the radii of the nodes represent the duration of fixations. Figure 1 presents scanpaths of four participants from our experiment for three sarcastic sentences *S1*, *S2* and *S3*. The x-axis of the graph represents the sequence of words a reader reads, and the y-axis represents a temporal sequence in which the fixations occur.

Consider a sarcastic text containing incongruous phrases *A* and *B*. Our general observation of scanpaths reveals the following: In most of the cases, a regression is observed when a reader starts reading *B* after skimming through *A*. On the other hand, if the reader slowly and carefully reads *A*, the fixation duration on *B* is significantly higher than the average fixation duration per word. A scanpath that does not exhibit one of the above mentioned properties indicates that the underlying incongruity (and hence, sarcasm) may not have been realized by the reader.

Figure 1 depicts two distinct cases through three example scanpaths, where differences in scanpaths are observed for *sarcasm_miss* and *sarcasm_hit*. The cases are explained below:

1. **Case 1. Lack of attention:** Consider sentence *S1*. When sarcasm is identified successfully, we see at least one regression from the phrase "original misconception" to the phrase "always cherish". On the other hand, participant $P_{Incorrect}$ have spent relatively smaller amount of time on these phrases. We do not observe any regressive saccade between the two incongruous phrases.
2. **Case 2. Lack of realization of underlying incongruity:** Consider sentences *S2* and *S3*. For sentence *S2*, participant $P_{Incorrect}$ focuses on the portion of the text containing positive sentiment. This indicates that the participant may have formed a bias towards the sentiment of the sentence being positive, thereby, not realizing the underlying incongruity. For sentence *S3*, participant $P_{Incorrect}$ spends more amount of time on the phrase "cheesy commercialism". The "negative intent" (and hence, the incongruity) of this phrase may not have been realized.

Though it is quite difficult to arrive at a conclusion regarding how sarcasm understandability is captured in scanpaths, our analysis gives us motivation to exploit properties from the eye-movement patterns to build systems for the task of sarcasm understandability prediction.

5.3 A Note on Complexity of Sarcasm

The complexity of sarcasm is not same across all texts. The underlying incongruity of sarcasm may either be **Explicit**, where a positive sentiment phrase is followed by a negative sentiment phrase, (as in "I love being ignored") or **Implicit**, where a positive sentiment phrase is followed by a *negative situation* that may not contain any sentiment bearing word (as in " It's an all star salute to Disney's cheesy commercialism") (Joshi, Sharma, and Bhattacharyya 2015). Moreover, the cognitive load associated with sarcasm processing

Category	Feature Name	Type	Intent
Textual Features	Interjections (IJT)	Integer	Count of interjections
	Punctuations (PUNC)	Real	Count of punctuation marks
	Discourse Connectors (DC)	Real	Count of discourse connectors
	Explicit Incongruity (EXP)	Integer	Number of times a word follows a word of opposite polarity
	Largest Pos/Neg Subsequence (LAR)	Integer	Length of the largest series of words with polarities unchanged
	Positive words (+VE)	Integer	Number of positive words
	Negative words (-VE)	Integer	Number of negative words
	Readability (RED)	Real	Flesch Readability Ease of the sentence (Kincaid et al. 1975)
	Number of Words (LEN)	Integer	Number of words in the sentence
Gaze Based Features	Avg. Fixation Duration (FDUR)	Real	Sum of fixation duration divided by word count
	Avg. Fixation Count (FC)	Real	Sum of fixation counts divided by word count
	Avg. Saccade Length (SL)	Real	Sum of saccade lengths (number of words travelled during saccade movements) divided by word count
	Regression Count (REG)	Real	Total number of gaze regressions
	Skip count (SKIP)	Real	Number of words skipped divided by total word count
	First part fixation duration (F1DUR)	Real	Average duration of fixation on the first half of the sentence.
	Second part fixation duration (F2DUR)	Real	Average duration of fixation on the second half of the sentence.
	Count of regressions from second half to first half of the sentence (RSF)	Real	Number of regressions from second half of the sentence to the first half of the sentence (given the sentence is divided into two equal half of words)
	Largest Regression Position (LREG)	Real	Ratio of the absolute position of the word from which a regression with the largest amplitude (number of pixels) is observed, to the total word count of sentence
	Scanpath Complexity (SPC)	Real	Product of Average fixation duration, Average saccade length and Average regression count

Table 3: Features for Sarcasm Understandability Prediction

depends on the degree of context incongruity between the statement and the context (Ivanko and Pexman 2003). Intuitively, scanpaths of readers also vary according to the type and the degree of incongruity in the text. Hence, a framework designed for sarcasm understandability prediction has to be aware of the type and degree of context incongruity in the sarcastic text, along with the eye-gaze attributes. Our feature design takes these aspects into consideration.

6 Predicting Sarcasm Understandability

Our framework for sarcasm understandability prediction is based on *supervised learning, i.e.*, it learns a predictive model on *training data* and a set of *features* (i.e. properties of each example of the training data). Our dataset contains 350 sarcastic sentences, each annotated by 7 participants. This amounts to a total of 2450 examples for which both eye-movement data and sarcasm understandability labels are available (discussed in section 4). From these, 48 instances are discarded due to poor quality of recorded eye-gaze patterns. Our feature-set comprises (i) **Textual features**, that capture the degree of incongruity of sarcastic text (taken from (Joshi, Sharma, and Bhattacharyya 2015; Riloff et al. 2013)) and (ii) **Gaze features**, that relate to sarcasm understandability (discussed in Section 5.1). Textual features are computed using in-house lexicons of interjections, discourse connectors, MPQA lexicon[6] and NLTK[7]. The feature-set is discussed in Table 3.

[6]http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[7]http://www.nltk.org/

6.1 Applying Multi-instance Classification

For each of the 350 sarcastic sentences, eye-movement data from multiple annotators are available. Instead of considering individual data for each sentence as a single example, we choose **Multi Instance Learning** for classification. Multi-instance (MI) learning differs from standard supervised learning in a way that each example is not just a single instance: examples are collections of instances, called **bags**. In our case, all of these bags correspond to one sentence and the instances of each bag correspond to gaze and textual features derived for each participant. Textual features are repeated for each instance in a bag. We apply **Multi-instance Logistic Regression (MILR)** (Xu and Frank 2004) implemented using the Weka API (Hall et al. 2009) under the *standard multi instance assumption*. After transforming our dataset to a multi-instance dataset, we performed a 5-fold cross validation[8]. Each fold has a train-test split of 80%-20% and each split contains examples from all seven participants. The train splits contain multiple instances per example whereas test splits contain one instance per example. This emulates a real life situation where our system would need to predict sarcasm understandability for any new sarcastic sentence for a new instance of reading.

In the absence of any existing system to compare our results with, we propose two baselines- (a) Baseline1: A classifier that generates predictions by respecting the training sets class distribution and (b) Baseline2: An MILR based classifier that considers the *average time taken by the reader*

[8]The system performs badly, as expected, in a Non-MI setting. The F-scores for SVM and Logistic Regression classifiers are as low as 30%. Hence, they are not reported here.

Class	sarcasm_miss			sarcasm_hit			Weighted Avg.			Kappa
	P	R	F	P	R	F	P	R	F	Avg.
	Baseline1: Classification based on class frequency									
All	16.1	15.5	15.7	86.5	87	86.7	85.9	86.71	86.3	0.014
	Baseline2: MILR Classifier considering time taken to read + textual features									
All	23.6	86.9	78.2	11.5	94.1	82.7	15.4	90.4	80	0.0707
	Our approach: MILR Classifier considering only gaze features									
All	82.6	36	50	89.9	98.7	94.1	88.8	89.4	87.5	0.4517
	Our approach: MILR Classifier considering gaze + textual features									
Quote	68.1	47.5	56.0	91.8	96.3	94.0	88.4	89.4	88.6	0.5016
Movie	42.9	36.6	39.5	88.6	91.0	89.8	81.4	82.5	81.9	0.293
Twitter	63.0	61.7	62.4	94.4	94.7	94.6	90.4	90.5	90.5	0.5695
All	**87.8**	**61**	**72**	**94.1**	**98.6**	**96.3**	**93.2**	**93.5**	**93**	**0.6845**

Table 4: Classification results for sarcasm understandability prediction. P→ Precision, R→ Recall, F→ F˙score, Kappa→ Kappa statistics showing *agreement of the predicted labels with the gold labels*

.

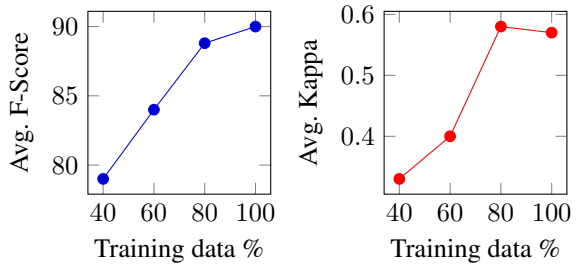

Figure 2: Effect of training data size on classification accuracy in terms of F-score (a) and *Kappa* statistics (b).

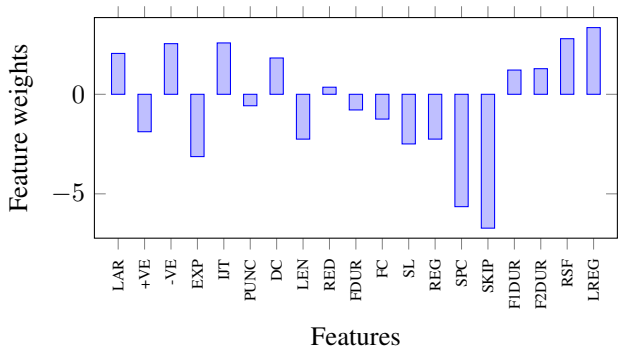

Figure 3: Significance of features as observed by the weights of the MILR classifier

along with other textual features for prediction. Please note that *average time taken by the reader* has not been taken as a feature in our best reported system due to its negative effect on the overall accuracy.

The results of our classification are shown in Table 4 for the whole data-set and individual domains (*viz.*, Quotes, Movie and Twitter). Our best system outperforms the baseline classifiers by considerable margins. Our system achieves an F-score of **72%** for *sarcasm_miss* class and an overall F-score of **93%**. The domain-wise performance is much less than that on the mixed domain data-set. It is difficult to explain why the performance of our system on certain domain is better/worse than others, given that systems for individual domains are left with small amounts of training data.

It is interesting to note that by considering the gaze features alone, the system still performs with reasonable accuracy, indicating that sarcasm understandability, the way we define it, is indeed an attribute of the reader. However, we notice that while precision is high (82.6%) for the *sarcasm_miss* class, the recall is quite low (36%). We speculate that, readers' eye-movement patterns may have been affected by other forms of linguistic complexities (like word difficulty, word sense ambiguity *etc.*). In the absence of the textual features (like readability) that help handle such complexities to some extent, some of the instances of *sarcasm_miss* are misclassified as *sarcasm_hit*, thereby, reducing the recall for *sarcasm_miss* class.

To see the impact of training data size on the classification results, we create a stratified random train-test split of 80%:20%. We train our classifier with 100%, 90%, 80% and 70% of the training data. The *F-scores* and *Kappa* statistics on the test set for various training data sizes are shown in Figure 2. The trend of F-score indicates that adding more data may increase the accuracy of our system.

We analyze the importance of our features by ranking them based on their weights learned by *Multi-instance Logistic Regression*. As shown in Figure 3, the negative and positive values of the features indicate their support towards *sarcasm_hit* and *sarcasm_miss* classes. The y-axis values show the weights of the features, indicating their predictive power. We observe that *average skip* has the maximum predictive power followed by *scanpath complexity*.

6.2 Error Analysis

The classification error could be attributed to a number of factors. First, our data-set is highly class imbalanced as cases of *sarcasm_miss* are significantly less than that of *sarcasm_hit* (with a class ratio of 1:8). This affects the learning of our classifier. Errors may have been introduced in feature extraction due to limitations of the NLP tools and errors

committed by the eye-tracking hardware.

7 Conclusion and Future Work

As far as we know, our work of predicting readers' sarcasm understandability is the first of its kind. We have tried to establish the relationship between sarcasm understandability and readers' eye movement patterns and proposed a predictive framework based on this observation. Our immediate future plan is to gather more training data, include more insightful features and explore additional techniques to address class-imbalance problem more efficiently. Since the motivation of this work comes from the increasing usage of eye-trackers in hand-held gadgets, we aim to check the usefulness of our technique on a much larger data-set, collected using mobile eye-trackers. Moreover, instead of considering sarcasm understandability as a two class problem, we plan to work towards building a framework that gives real valued scores for sarcasm understandability, indicating to what extent sarcasm has been understood. Our overall observation is that cognition cognizant techniques involving eye-tracking look promising for sarcasm understandability prediction.

Acknowledgment: We thank the members of CFILT Lab and the students of IIT Bombay for their help and support.

References

Barbieri, F.; Saggion, H.; and Ronzano, F. 2014. Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th WASSA, Association for Computational Linguistics*.

Camblin, C. C.; Gordon, P. C.; and Swaab, T. Y. 2007. The interplay of discourse congruence and lexical association during sentence processing: Evidence from {ERPs} and eye tracking. *Journal of Memory and Language* 56(1):103 – 128.

Campbell, J. D., and Katz, A. N. 2012. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes* 49(6):459–480.

Carvalho, P.; Sarmento, L.; Silva, M. J.; and De Oliveira, E. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, 53–56. ACM.

Clark, H. H., and Gerrig, R. J. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General* 113(1):121–126.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth CoNLL*. Association for Computational Linguistics.

Filik, Ruth; Leuthold, H. W. K. P. J. 2014. Testing theories of irony processing using eye-tracking and erps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Gibbs, R. W. 1986. Comprehension and memory for nonliteral utterances: The problem of sarcastic indirect requests. *Acta Psychologica* 62(1):41 – 57.

Giora, R. 1995. On irony and negation. *Discourse processes* 19(2):239–264.

González-Ibánez, R.; Muresan, S.; and Wacholder, N. 2011. 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-Volume 2*, 581–586. Association for Computational Linguistics.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*.

Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; and Van de Weijer, J. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Ivanko, S. L., and Pexman, P. M. 2003. Context incongruity and irony processing. *Discourse Processes* 35(3):241–279.

Jorgensen, J.; Miller, G. A.; and Sperber, D. 1984. Test of the mention theory of irony. *Journal of Experimental Psychology: General* 113(1):112.

Joshi, A.; Mishra, A.; Senthamilselvan, N.; and Bhattacharyya, P. 2014. Measuring sentiment annotation complexity of text. In *Association of Computational Linguistics (Daniel Marcu 22 June 2014 to 27 June 2014)*. Association for Computational Linguistics.

Joshi, A.; Sharma, V.; and Bhattacharyya, P. 2015. Harnessing context incongruity for sarcasm detection. *Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China* 757.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Kutas, M., and Hillyard, S. A. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207(4427):203–205.

Liebrecht, C.; Kunneman, F.; and van den Bosch, A. 2013. The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013* 29.

Malsburg, T., and Vasishth, S. 2011. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language* 65(2):109–127.

Martınez-Gómez, P., and Aizawa, A. 2013. Diagnosing causes of reading difficulty using bayesian networks.

Maynard, D., and Greenwood, M. A. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of LREC*.

Mishra, A.; Bhattacharyya, P.; and Carl, M. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Conference of Association for Computational Linguistics (ACL), Sofia, Bulgaria*.

Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*.

Parasuraman, R., and Rizzo, M. 2006. *Neuroergonomics: The brain at work*. Oxford University Press.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124(3):372.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of Empirical Methods in Natural Language Processing*, 704–714.

Shamay, S.; Tomer, R.; and Aharon, J. 2005. The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology* 19(3):288.

Xu, X., and Frank, E. 2004. Logistic regression and boosting for labeled bags of instances. In *Advances in knowledge discovery and data mining*. Springer. 272–281.