Industrial Management & Data Systems

Detecting sarcasm in customer tweets: an NLP based approach
Shubhadeep Mukherjee, Pradip Kumar Bala,

The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/0263-5577.htm

Detecting sarcasm in customer tweets: an NLP based approach

Shubhadeep Mukherjee
*Department of Information Systems,
Indian Institute of Management Ranchi, Ranchi, India, and*
Pradip Kumar Bala
Operations Management and Decision Science, XIMB, Bhubaneswar, India

Received 11 June 2016
Revised 5 October 2016
Accepted 15 November 2016

Abstract

Purpose – The purpose of this paper is to study sarcasm in online text – specifically on twitter – to better understand customer opinions about social issues, products, services, etc. This can be immensely helpful in reducing incorrect classification of consumer sentiment toward issues, products and services.

Design/methodology/approach – In this study, 5,000 tweets were downloaded and analyzed. Relevant features were extracted and supervised learning algorithms were applied to identify the best differentiating features between a sarcastic and non-sarcastic sentence.

Findings – The results using two different classification algorithms, namely, Naïve Bayes and maximum entropy show that function words and content words together are most effective in identifying sarcasm in tweets. The most differentiating features between a sarcastic and a non-sarcastic tweet were identified.

Practical implications – Understanding the use of sarcasm in tweets let companies do better sentiment analysis and product recommendations for users. This could help businesses attract new customers and retain the old ones resulting in better customer management.

Originality/value – This paper uses novel features to identify sarcasm in online text which is one of the most challenging problems in natural language processing. To the authors' knowledge, this is the first study on sarcasm detection from a customer management perspective.

Keywords Text mining, Natural language processing, Artificial intelligence, Data mining, Business intelligence, Sarcasm detection

Paper type Research paper

1. Introduction

With the arrival of the information age, social media has become one of the most powerful tools for businesses to identify customer attitudes toward their products and services. Modern businesses are becoming increasingly dependent on the online medium to attract customers and reduce the customer churn rate (Kacen *et al.*, 2013). Various data mining techniques are applied by organizations to understand customer preferences and opinions.

One of the most popular techniques for analyzing online data is sentiment analysis. In sentiment analysis, opinions can be classified into positive, neutral or negative (Pang and Lee, 2008). Sentiment analysis about products can help in determining customer preferences and dislikes. In spite of its success, under certain circumstances sentiment analysis can be gravely inadequate. One such situation is when the sentences are laden with sarcasm, for example, sarcastic user tweets. It is quite possible that a sarcastic tweet, which mockingly praises a product while actually deriding it, be classified as positive customer emotion. Sarcasm, being a special type of communication, where the explicit meaning differs from the implicit one, cannot be effectively identified with conventional data mining techniques such as sentiment analysis (Yee Liau and Pei Tan, 2014).

Macmillan English Dictionary defines sarcasm as the activity of saying or writing the opposite of what one means or of saying in a way intended to make someone else feel stupid or show them that one is angry (Rundell and Fox, 2002). With sophistication of language, the use of sarcasm in verbal and written text has become the norm. However, automatic detection of sarcasm is still in its infancy. The ambiguous nature of sarcasm makes it

0263-5577
DOI 10.1108/IMDS-06-2016-0207

difficult even for humans to detect it in sentences. Despite the difficulties, the huge benefit of detecting sarcasm has been recognized in many computer interaction-based applications, such as review summarization, dialogue systems and review ranking systems (Davidov *et al.*, 2010). From a business perspective, detecting sarcasm can be crucial in correctly categorizing customer opinions about products, services and social issues, all of which suffer from a high threat of being incorrectly categorized.

This makes sarcasm detection from unstructured text data a relevant and challenging problem. This is also because it is unaided by any visual or vocal cues that assist humans in understanding sarcasm. One of the major issues in sarcasm detection is the absence of naturally occurring expressions that can be used for training purposes (Davidov *et al.*, 2010). In the case of microblogs, such as Twitter, messages can be annotated with hashtags that are an indication of the sentiment being expressed in tweets. These hashtags are reliable indicators of the emotion being expressed by the tweets, as the author explicitly conveys the emotion of the tweet through them (e.g. #happy, #joy, #sad). We utilized this behavior to formulate hashtags (#sarcasm, #sarcastic) for our data set. We considered the sentences that end in #sarcasm or #sarcastic to be the gold standard for sarcastic sentences. We did supervised learning, using Naïve Bayes and maximum entropy classifier to differentiate between a sarcastic and a non-sarcastic tweet. In our knowledge, this is the first attempt to study and understand sarcasm from a customer management perspective.

We trained our classifiers on multiple different feature types. The feature set, we emphasize on, consists of function words, part of speech tags, part of speech *n*-grams and their various combinations. At first, we have used topic as well as writing style-based features to classify the tweets for sarcasm detection. We did not come across any work in sarcasm detection literature which has tried to capture authorial style-based features. Our work thus adds a new dimension to natural language processing (NLP)-based research on sarcasm detection.

An English sentence can be broadly said to consist of two types of words – function words and content words. Function words are the words that have little or no significant meaning outside the premise of the sentence. On the other hand, content words are the words that have meaning even outside the context of the sentence (William Collins Sons & Co. Ltd, 2009). Examples of function words are – the, and, he, not, etc. Examples of content words are – school, dog, angry, etc. If we were to consider an English sentence in its entirety, it would consist of these two categories of words.

Extant literature states that the authorial or writing style is best captured by the function words and the part of speech used in the sentence (Argamon *et al.*, 2003). Koppel (2002) state that categorization by topic is typically based on keywords that reflect a document's content whereas categorization by author style uses precisely those features that are independent of context. Authorial style-based classification has been applied successfully in gender classification of regular text as well as microblogs (Argamon *et al.*, 2003; Mukherjee and Bala, 2016). We propose that the content of the tweets as well as the authorial or writing style both contribute to the sarcasm present in the tweets. We have used features that are independent of the content of the text in conjunction with other topic or content-based features. By content-based features, we mean those features which are an integral part of the text and give the text its meaning. For example, if we consider a tweet "Amazon, love your customer service, really amazing!" then the content words are – "Amazon," "love," "customer," "service" and "amazing." The rest of the words – "your" and "really," are function words or writing style-based features that can vary from author to author.

We hypothesize that sarcasm in a sentence is dependent not only on the content words of the sentence but also on the authorial or writing style of the author, which are best depicted by function words and parts of speech of the sentences (Argamon *et al.*, 2003; Koppel, 2002).

We downloaded multiple tweets by various twitter users on a range of products/services and differentiated the sarcastic tweets from the non-sarcastic ones.

Before delving further, we would like to discuss some of the various sarcastic tweets that we have come across in our tweet data sets:

(1) Apple pencil is a college humor video!!

(2) I'd like to thank Michele Obama for making the fruit snacks in the lunch room 90 percent; tinier! Really changed my whole life with that one.

(3) Apple's sports watch will lead millions of nerds working really hard for like 3 days.

(4) The phone battery doesn't even last an hour after charging to a 100 percent, Thank you Samsung!

All the above tweets convey sarcasm. The first tweet mocks the new Apple product, the Apple pencil by comparing it with a College Humor video, College Humor, as the name suggests, is a website with satirical takes on day-to-day events. The second tweet makes a mockery of the fight against obesity project started by Michelle Obama, the First Lady of the USA. The third tweet mocks the diehard Apple product fans of having to work really hard to get the product "Apple sports watch," which we understand does not actually require physical hard work on part of the customer. The fourth tweet is used to lament the fact that the Samsung mobile, that the user owns, has very low battery power and does not even last an hour, the author mockingly thanks Samsung, which we understand from the context is insincere.

The customer tweet examples mentioned above appear to be neutral or positive; however, as can be observed from the context we provide, they are either mocking a product or a policy. These tweets have a high probability of being classified as expressing positive or neutral emotion whereas in reality they convey negative emotions using sarcasm as a tool.

The objective of this research work is twofold. The first one is to identify characteristics in sentences which can be used for detecting sarcasm in tweets. This has been done by applying multiple different feature types consisting of both content as well as writing style-based features with an emphasis on the writing style-based features. This bridges an important gap in the NLP literature on sarcasm detection as there are no studies till date on sarcasm detection based on writing/authorial style. Second, this is the first study on sarcasm detection from a customer management perspective. We have done this by considering the tweets of customers about various products and services and identifying whether the tweets are sincere or sarcastic.

We have used a data set of 5,000 tweets consisting of sarcastic and non-sarcastic tweets and applied feature types that capture the authorial or writing style across five data sets. Our experiments reveal that a combination of function words and content words of the tweet are ideal for detecting sarcasm.

The rest of this paper is divided into four sections. Section 2 discusses the literature on sarcasm detection and the limitations of sentiment analysis. Section 3 presents the data collection and the research methodology. Section 4 deals with the classification results obtained through the two supervised learning algorithms using accuracy and *F*-measure. The section also identifies the most differentiating authorial style-based features between a sarcastic and a non-sarcastic tweet. Section 5 concludes the study with some discussions on the utility of this research work. The section also discusses some of the limitations of this study.

2. Literature review

We have divided the literature review section into three parts. In the first part, we give a generalized overview of the research conducted in understanding sarcasm and its use across areas. In the second part, we provide a more specific review of the work done in automatic sarcasm detection, which is the premise of our research. In the third part, we review the current literature on sentiment analysis and discuss its limitations.

2.1 Overview

Sarcasm is a form of speech act in which the speakers convey their message in an implicit way (Davidov *et al.*, 2010). The implicitness of the statements makes it hard for humans to decide whether a statement is sarcastic or not. Sarcasm has been studied in depth in linguistics, psychology and cognitive sciences (Gibbs, 1986; Gibbs and Colston, 2007; Kreuz and Glucksberg, 1989; Utsumi, 2002). Jorgensen (1996) stated that sarcasm arises from figurative meaning as opposed to literal meaning. Clark and Gerrig (1984) proposed that sarcasm cancels the indirectly negated message by replacing it with the implicated one. Giora (1995) refuted the claims of the earlier researchers by stating sarcasm to be a mode of indirect negation which requires processing of both the negated and implicated messages. Later *et al.* studied the inherent complexity of sarcasm and its effect on sarcasm processing time.

Table I summarizes the major works in sarcasm.

2.2 Automatic sarcasm detection of online text

One of the remarkable works on sarcasm detection in the field of text mining has been done by Tsur and Davidov (2010). The authors used a semi-supervised algorithm for sarcasm detection (SASI) in product reviews. It consisted of two stages: semi-supervised pattern acquisition and sarcasm classification. They used Amazon review for books and products for the task. The pattern acquisition task in their work consisted of pattern extraction, selection and matching. Additionally, they also used punctuation-based features for classification. Davidov *et al.* (2010) used the semi-supervised learning based on SASI to classify tweets and amazon product reviews. González-Ibánez *et al.* (2011) used basic supervised learning techniques to classify tweets for sarcasm using hashtags (#sarcasm) as the gold standard. The authors used both lexical and pragmatic features for the classification job. They also did a comparative study of the human performance with the machine learning algorithm on accuracy of sarcasm detection. The authors also identified most discriminating features using multiple classes to gain insights in the problem. Most recently, Justo *et al.* (2014) have used a range of different features, such as unigrams to classify tweets for sarcasm.

Table II summarizes the work done in the field of automatic sarcasm detection.

Author(s)	Results/findings
Jorgensen	Sarcasm arises from figurative meaning as opposed to literal meaning (Jorgensen, 1996)
Gibbs	Ease of processing and memory for sarcastic utterances depends crucially on how explicitly a speaker's statement echoes either the addressee or some other source's putative beliefs, opinions, or previous statement (Gibbs, 1986)
Kreuz and Glucksberg	Listeners recognize sarcasm when they perceive that a speaker is alluding to some antecedent state of affairs (Kreuz and Glucksberg, 1989)
Giora	Sarcasm is a mode of indirect negation which requires processing of both the negated and implicated messages (Giora, 1995)
Utsumi	Cognitive model of how poetic effects are achieved by a work of literature, especially by individual figurative expressions such as metaphor and irony (Utsumi, 2002)
Ivanko and Pexman	Direct action model of figurative language processing (Ivanko and Pexman, 2003)
Gibbs and Colston	Irony in language and thought (Gibbs and Colston, 2007)

Table I.
General literature review on study of sarcasm

Author(s)	Results/Findings
Tsur and Davidov	Motivation for using sarcasm in online communities and social networks (Tsur and Davidov, 2010)
Davidov *et al.*	Dependencies and overlap between different sentiment types represented by smileys and Twitter hashtags (Davidov *et al.*, 2010)
González-Ibánez *et al.*	Impact of lexical and pragmatic factors on machine learning effectiveness for identifying sarcastic utterances (González-ibáñez *et al.*, 2011)
Kunneman, Liebrecht, and van den Bosch	Sarcasm is signaled by hyperbole, using intensifiers and exclamations; in contrast, non-hyperbolic sarcastic messages often receive an explicit marker (Kunneman *et al.*, 2014)
Maynard and Greenwood	Impact of sarcasm on sentiment analysis (Maynard and Greenwood, 2014)
Justo *et al.*	Sarcasm detection task benefits from the inclusion of linguistic and semantic information sources (Justo *et al.*, 2014)

Table II. Literature review on automatic sarcasm detection

2.3 Current limitations of sentiment analysis

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude toward a particular topic, product, etc., is positive, negative or neutral (Turney, 2002). Research in sentiment analysis took shape in the early 2000s. Since then a lot of progress has been made in the field. We critique some of the recent research on sentiment analysis.

In 2012, Saif *et al.* used semantic features along with extracted entities to carry out sentiment analysis on twitter. The authors showed that semantic features produce better results than sentiment – bearing topic analysis. Ghiassi *et al.* (2013) did twitter brand sentiment analysis using supervised feature reduction using *n*-grams and statistical analysis. The authors successfully identified mildly positive and mildly negative emotions in neutral comments using different *n*-grams as features. More recently, Khan *et al.* (2015) combined lexicon- and learning-based methods for twitter sentiment analysis. The authors present a novel lexicon-based approach to perform an entity-level sentiment analysis followed by automatic identification of opinionated tweets, finally training a classifier to assign polarities to the tweets.

The above-mentioned recent studies have effectively contributed toward the development of sentiment analysis literature by using novel techniques and some unique feature types, however, there remain open problems in sentiment analysis which are yet to be addressed satisfactorily, such as automatic systems for noise removal, a universal opinion grading system or sarcasm detection in sentences (Martinez-Camara *et al.*, 2014; Ravi and Ravi, 2015). One of the major problem areas in sentiment analysis literature is sarcasm or irony detection in sentences. Ravi and Ravi (2015), in their detailed review of opinion analysis, mention that little or no studies have been devoted to study of irony or sarcasm detection, making it an open problem in the area. They further point out the need to develop computational approaches for detecting sarcasm using the appraisal theory. According to the appraisal theory, our interpretation of a situation causes an emotional response that is based on that interpretation (Scherer, 2001). Our current approach addresses the above-mentioned research gap by using authorial style-based features for sarcasm detection in customer tweets. We would like to emphasize that none of the above-mentioned studies have taken into account authorial or writing style-based features for sentiment or sarcasm detection. This is a unique contribution to the field of opinion mining as well as sarcasm detection.

Psychological research on sarcasm has revealed that use of sarcasm is associated with socio-economic class and profession (Katz *et al.*, 2004). Also, a person's profession, habits and social circle affect the way he/she thinks, talks and writes (Gergen, 1999). Both these facts combined reveal that the use of sarcasm is related to the writing style of the author. For instance, a comedian would write quite differently from to a school teacher. Two students in the same class could write in very different styles on the same given topic, based on the differences in their ability to imagine and articulate. Hence, for the study of sarcasm in written text, the study of writing/authorial style becomes crucial. The significance of writing style has already been observed in other problems of NLP, such as gender detection (Argamon *et al.*, 2009; Mukherjee and Liu, 2010).

In this paper, we have proposed that for effective identification of sarcasm, both the content as well as writing style of the author plays a crucial role. Through our classification algorithms, we have identified a set of features that capture authorial style. We got an accuracy of more than 70 percent with the features proposed by our study, which is on the higher side in sarcasm detection literature. It must be noted that we have worked with a reasonable data set (5,000 tweets) size and increasing the data set size could lead to further improvement in the accuracy levels obtained. We have also reported satisfactory levels of precision, recall and *F*-measure as observed through our experiments.

3. Methodology

Twitter is a microblogging service that allows users to post short updates limited to 140 characters. Research shows that more than 80 percent of twitter users update their status on a daily basis, making it a reliable source of research data (Thelwall *et al.*, 2011). Due to its inherent heterogeneity, twitter has become an important source of online opinion.

We have divided this section into multiple subsections. We first describe the data preprocessing part and then move onto the specifics, such as *K*-fold cross-validation and feature extraction. Finally, we describe the classification algorithms used in the research.

3.1 Data preprocessing

We downloaded around 15,000 tweets using hashtags such as #sarcasm, #sarcastic along with sincere tweets using R software. Hashtags have been found to be useful and have been extensively used in earlier research on twitter (Tsur and Davidov, 2010). Retweets as well as tweets from the same twitter handle were removed. Also, around 300 tweets which could not be satisfactorily classified as sarcastic or non-sarcastic were removed from the data set. We ended up with around 5,000 tweets with 2,600 sarcastic and 2,400 non-sarcastic tweets. To train and test the classifiers, the data were split into two sets randomly. The data set was divided into a ratio of 3:1. The mentioned ratio has been extensively applied in classification literature (Schürer and Muskal, 2013). A tenfold cross-validation was performed on the training set. In choosing the training testing ratio, the stress is on generalizability of the results which is achieved by the *K*-fold cross-validation as explained later in this section (Domingos, 2012) (Figure 1).

3.2 K-fold cross-validation

One needs to ensure that the training data does not overfit the training set as it could drastically distort the result for the test set. This is usually addressed by the *K*-fold cross-validation. In our case, we have taken $K = 10$ which is the usual norm in classification data training (Pennacchiotti and Popescu, 2011). A tenfold cross-validation entails dividing the data set into ten equal random folds and nine of them are used for training and one for testing or validation. The whole process is repeated ten times with each of the sub-folds being used for validation exactly once. This ensures that the model generalizes to an independent data set and does not overfit (Kohavi, 1995).

Various features (mentioned later in this section) were extracted from tweets and selected from the training set. The features were then tested for accuracy and *F*-measure on

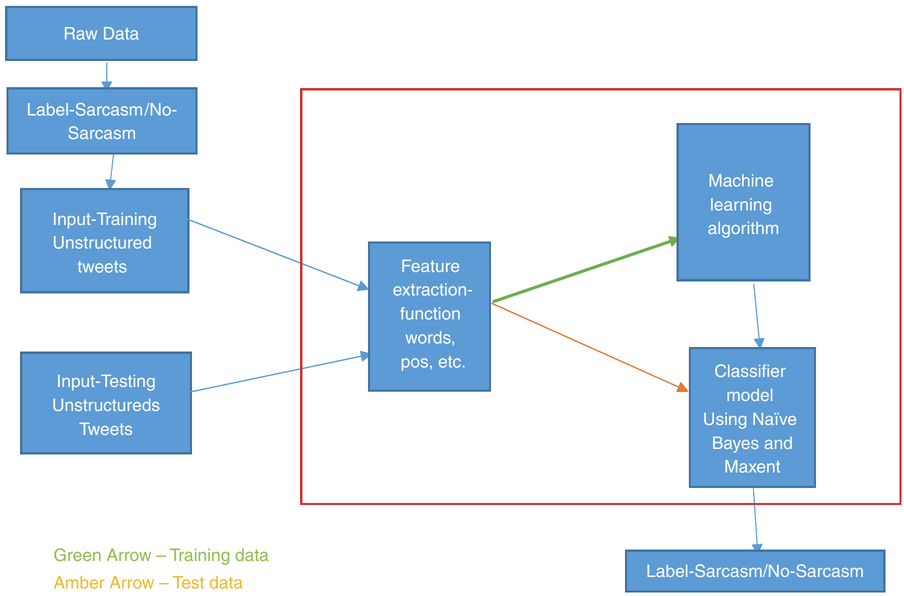


Figure 1.
Sarcasm detection in tweets based on supervised learning

the test set. We started with a small number of tweets and progressively increased the number to observe its effect on the classification accuracy and the *F*-measure. One must bear in mind that, for a small data set, the method for manual cleaning and labeling tweets is standard in supervised learning. We have emphasized on extracting features that have not been used in extant literature.

We now explain the feature extraction method used in our work.

3.3 Feature extraction

Feature extraction is a method to reduce the amount of resources required to describe a data set (Guyon and Elisseeff, 2003). When analyzing complex data, one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power. Also, this makes the classification algorithm overfit the training sample and generalize poorly to new samples. Hence, feature extraction becomes essential while dealing with classification problems with large number of variables.

We have used a comprehensive list of features for the purpose of classification. Using multiple features gives us the opportunity to compare the different accuracies and *F*-measures achieved for different features as well as a combination of features. The features we have used for our study are as follows:

(1) Content words – the *Collins Dictionary* defines content words as words to which an independent meaning can be given by reference to a world outside any sentence in which the word may occur (Winkler, 2012).

(2) Function words – function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. According to the *Collins English Dictionary*, function words are words, such as "the," with a particular grammatical role but little identifiable meaning (Klammer *et al.*, 2000).

(3) Part of speech tags – it is the process of marking up a word in a text with reference to a corpus as corresponding to a particular part of speech, based on both its definition, as well as its context. We have used part of speech tags as features for our training set (e.g. This/PNN, is/VB, a/ART, dog/NN) (Church, 1989). Here, PNN means pronoun, VB means verb, ART means article and NN means a noun. Any English sentence could be broken down into its part of speech tags.

(4) Part of speech n-grams – an n-gram model is a type of probabilistic language model for predicting the next item in a sequence in the form of a $n-1$-order Markov model. The prediction could be done on the basis of a single preceding item (bigram), two preceding items (trigram) or more items (four gram, five gram, etc.). In our case, the items are part of speech of the words used in the sentences. We have used trigrams of part of speech as features for our model (Koppel, 2002). Taking higher n-grams, such as four or five, have not been very effective in increasing classification efficiency in the past.

(5) Content words + function words – content words and function words have been extracted separately as different feature types. We have also extracted both these feature types. together and used as a single feature type to capture both style- and topic-based features.

(6) Function words + part of speech n-grams – here we have combined function words and part of speech n-grams and used them as a single feature for classification. These features exclusively capture style-based features.

(7) Content words + function words + part of speech n-grams – here we have combined the most informative content words, the function words and the part of speech n-grams as features for classifying the tweets. Using this feature type we have tried to capture both style-based as well as topic-based features.

After extracting the above-mentioned features we have applied two different classification algorithms, namely, the Naïve Bayes and the maximum entropy to classify the tweets. The rationale behind using the specific classifiers has been explained later in Subsection 3.4.

3.4 Classification method

Classification models in NLP are broadly of two types – generative and discriminative. Generative classifiers learn the joint probability of the inputs and the labels (classes like in our case sarcasm/non-sarcasm), and make the prediction by using the Bayes rule to select the most likely label. The discriminative classifiers model the posterior probability directly or learn a direct map of inputs to the class label (Ng and Jordan, 2002). Both types of classifiers have been used by researchers in the past. Yan and Yan (2006) used a generative classifier (Naïve Bayes) and Rao *et al.* (2010) used a discriminative classifier (support vector machine). We have formulated both the types of classification models, the Naïve Bayes model (generative classifier) and the maximum entropy model (discriminative classifier).

3.4.1 Naïve Bayesian Classifier. The Naïve Bayes classifier is a popular classification algorithm used extensively in document classification (Argamon *et al.*, 2007; Mukherjee and Liu, 2010). We have shown how the Naïve Bayes classifier works in the case of text classification. We considered a document vector model (Manning and Schutze, 1999) for representing a document with the help of terms which can be used as inputs.

Let us consider a tweet with some features of our interest $T=(F_1, F_2, \ldots, F_n)$. Here F can be any of the features based on which we would like to classify the tweets (content words, n-gram part of speech tags, function words, etc.). Given the tweet "T" we would like to predict whether it belongs to a particular category, namely, sarcastic or non-sarcastic.

Using Bayes' theorem we can write:

$$p(C/F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n/C)}{p(F_1, \ldots, F_n)} \tag{1}$$

where C = {sarcastic, non-sarcastic}. F_i represents the features selected as inputs for developing the classification model as per Table I.

The Naïve Bayes assumption for a classification task is as follows:

$$p(F_1, \ldots, F_n|C) = p(F_1|C)p(F_2|C), \ldots, p(F_n|C). \tag{2}$$

The assumption of independence between or amongst the features is considered in the above expression. In the case of tweets, it will mean that the two words (a feature) in a tweet occur independent of each other. Although the assumption is simplistic it has been shown to work well in earlier research (Yan and Yan, 2006).

This equation could now be written as:

$$p(C/T) = \frac{p(C)p(F_1|C)p(F_2|C), \ldots, p(F_n|C)}{p(T)}. \tag{3}$$

We then compute the ratios of the posterior probabilities P (C = sarcastic/T)*and* P (C = non-sarcastic/T) of the two classes for a given document. This is done by calculating the prior probabilities $p(C)$ and the conditional probabilities of $P(Fi|T)$. The tweet is then classified to the class that yields the higher probability.

3.4.2 Maximum entropy classifier. Unlike the Naïve Bayes classifier, the maximum entropy classifier does not assume that the features are conditionally independent of each other. Maximum entropy is therefore a less restrictive model than Naïve Bayesian model. It is based on the principle of maximum entropy and from all the models which fit the training data, it selects the one which has the highest entropy. The maximum entropy classifier requires more time to train compared to Naïve Bayes due to the optimization problem that needs to be solved in order to estimate the parameters of the model.

We construct a stochastic model which accurately represents the behavior of the process. We take the contextual information a as input (function words, unigram, bigram, etc.) of a document and produce the output value b.

The initial step of constructing this model is to collect training data that consists of samples represented in the following format: (a_i, b_i) where the a_i includes the contextual information of the document and b_i its class. The next step is to summarize the training sample in terms of its probability distribution:

$$p(a, b) = \frac{1}{N} \times \text{ number of times that } (a, b) \text{ occurs in the sample set} \tag{4}$$

where N is the size of the training set.

We use the above empirical probability distribution in order to construct the statistical model of the random process which assigns texts to a particular class by taking into account their contextual information.

We use the following function:

$$f_j(a, b) = \begin{cases} 1 & \text{if } b = L_i \text{ and } a \text{ contains } K_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where f_j is the feature function that returns 1 when the class of the function is L_i and the document contains the word K_i. We express any statistic of the training data set as the expected value of the appropriate binary-valued indicator function f_j.

The expected value of f_j with respect to the distribution p (a, b) is:

$$p(f_j) \equiv \sum_{a,b} p(a,b) f_j(a,b). \tag{6}$$

If each training sample (a, b) occurs once in training data set then p (a, b) is equal to $1/N$.

We constrain the expected value that the model assigns to the expected value of the feature function f_j. The expected value of feature f_j with respect to the model $p(b/a)$ is equal to:

$$p(f_j) \equiv \sum_{a,b} p(a) p\left(\frac{b}{a}\right) f_j(a,b) \tag{7}$$

where p (a) is the empirical distribution of a in the training data set and it is usually set equal to $1/N$.

By constraining the expected value to be equal to the empirical value:

$$\sum_{a,b} p\,(a,b) f_j(a,b) = \sum_{a,b} p\,(a) p\left(\frac{b}{a}\right) f_j(a,b) \tag{8}$$

Equation (8) is the constrain equation which depends on the number of feature functions.

The constraint in Equation (8) can be satisfied by multiple models; however, according to the principle of maximum entropy, the model should be the most uniform amongst the ones that satisfy the constraint. One could also say that the model should have the maximum entropy to be selected:

$$p_{max} = \arg\ max_{p \in L} \left(-\sum_{a,b} p(a)\, p\left(\frac{b}{a}\right) \log\ \ p\left(\frac{b}{a}\right) \right) \tag{9}$$

It now becomes an optimization problem with Equation (8) as the constraint.

Both the above-mentioned techniques address the classification problem considering the classification boundary to be linearly separable. This gives satisfactory result in our case. The research could be extended to non-linear classifiers, such as K-nearest neighbors and support vector machine (Rao *et al.*, 2010).

4. Result and discussion

We now report the sarcasm detection ability of the classification algorithms based on the various feature types. The performance of these systems was measured by a variety of metrics, such as precision, recall, accuracy and *F*-measure. Accuracy is the percentage of instances predicted in the correct classes in a classification problem. However, in case of unbalanced classes, accuracy can give spurious results, in such cases *F*-measure in classification is a better metric. It is a measure that combines precision and recall by calculating the harmonic mean of precision and recall.

It is denoted in its common form by the following formula:

$$F - \text{measure} = \frac{2(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \quad (10)$$

where precision is the number of retrieved instances which are relevant and recall is the number of relevant instances which have been retrieved.

The confusion matrix is illustrated in Table III.

F-measure has been used in previous studies in classification literature as an overall assessment of performance of a classifier as it takes into account both precision and recall (Justo *et al.*, 2014). When measured by these metrics, each algorithm demonstrates its sarcasm detection capability. Once the data are trained on the training data set, both the maximum entropy algorithm and the Naïve Bayes algorithms are run on the test set. We used several tweet data sets with increasing number of tweets in each of them to identify the best feature type. We found that "content words and function words" give the best results for both the measurement metrics across the data sets. This is in line with our initial claim that features pertaining to authorial style as well as topic are crucial for classifying tweets.

Tables IV and V summarize the accuracy, *F*-measure and the best feature type for each of the tweet data sets. The rows that are made italics give the best results across the data sets and the feature types.

Figures 2 and 3 show the change in accuracy and *F*-measure as the number of tweets are increased across the data sets.

It can be observed from Figure 2 that there is a gradual increase in accuracy levels as the data set size increases. This consistent increase is a sign that the features used perform satisfactorily across the data sets. We have worked with a data set of

Confusion matrix	(Predicted class) Yes	No
(Actual class) Yes	TP	FN
No	FP	TN

Notes: Precision = TP/(TP + FP), recall = TP/(TP + FN), accuracy = (TP + TN)/(TP + TN + FP + FN), where TP, true positive; TN, true negative; FP, false positive; FN, false negative

Table III. Confusion matrix

No. of tweets	Best feature	Classifier	Accuracy
1,000	ContW	Naïve Bayes	0.57
2,000	ContW + FuncW	Naïve Bayes	0.6
3,000	ContW + FuncW	Naïve Bayes	0.63
4,000	ContW + FuncW	Naïve Bayes	0.67
5,000	*ContW + FuncW*	*Naïve Bayes*	*0.73*

Table IV. Best feature types and classifier across data sets (accuracy)

No. of tweets	Best feature	Classifier	*F*-measure
1,000	ContW	NaiveBayes	0.71
2,000	FuncW	NaiveBayes	0.74
3,000	ContW + FuncW	NaiveBayes	0.74
4,000	FuncW	Maximum Entropy	0.75
5,000	*ContW + FuncW*	*NaiveBayes*	*0.76*

Table V. Best feature types and classifiers across data sets (*F*-measure)

5,000 tweets only. The above trends show that if the data set size is increased the accuracy could improve further.

In Figure 3, we look at the *F*-measure trends across the data set.

We can observe that there has been an increase in the *F*-measure statistic as the data set size increases. This is consistent with the accuracy trends obtained earlier in Figure 2. We can safely say that the increase in accuracy of the model is not at the cost of the precision or recall but is due to the overall improvement in model as the number of relevant features increase with increase in the number of tweets.

In Figures 4-6, we compare accuracy with features across the data sets and the classifiers.

In Figures 4-6, we observe that "content words and function words" are the best features across the data sets and the classifiers. It can also be observed that Naïve Bayes classifier performs better than maximum entropy classifier for the authoritative data set (5,000 tweets).

Tables VI and VII summarize the differentiating capability of the authorial style-based features part of speech – *n*-grams and function words used in our model, respectively. The ratios show the likelihood of the expressions to be used in a sarcastic to a non-sarcastic sentence.

Table VI has four columns. The first column illustrates the part of speech *n*-gram used in the data set, second column provides information about the general usage of a part of speech *n*-gram in a tweet (sarcastic or non-sarcastic), the third column gives the likelihood ratio of the part of speech *n*-gram, of being used in a sarcastic or non-sarcastic sentence, and the fourth column provides the meaning of the particular part of speech *n*-gram.

To cite an example, the verb form BEM or "to be" in the present tense, first person singular, is twice more likely to be used in a non-sarcastic sentence than in a sarcastic sentence.

Next, we have a look at the top function words in our data set.

Table VI indicates that certain parts of speeches are more likely to be used in sarcastic tweets, namely, non-sarcastic ones. For instance, pronouns along with nominative verbs are almost six times more likely to be used in sarcastic tweets by customers than in

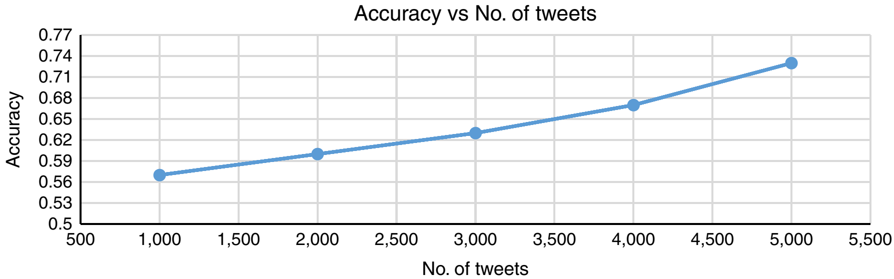


Figure 2.
Accuracy trends across the tweet data sets

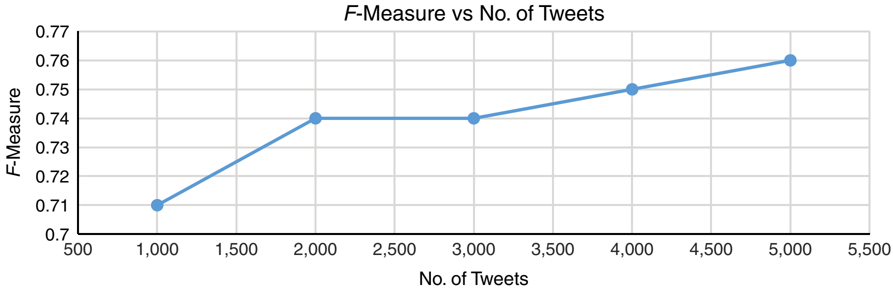
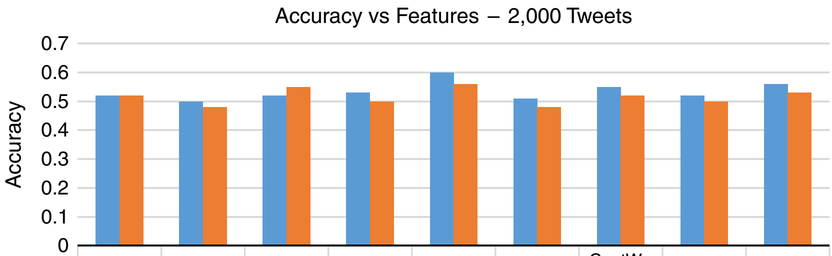


Figure 3.
F-measure trends across the tweet data sets

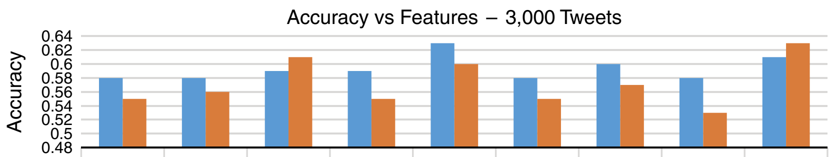


	AllWords	Pos tags	FuncW	ContW	ContW+FuncW	ngrampos	ContW+FW+ngram	ngrampos+FuncW	ngrampos+ContW
Accuracy-NV	0.52	0.5	0.52	0.53	0.6	0.51	0.55	0.52	0.56
Accuracy-ME	0.52	0.48	0.55	0.5	0.56	0.48	0.52	0.5	0.53

Features

Notes: NV, Naïve Bayes; ME, maximum entropy; Postags, part of speech tags; FuncW, function words; ContW, content words; ngrampos, part of speech n-grams

Figure 4.
Accuracy vs features – 2,000 tweets

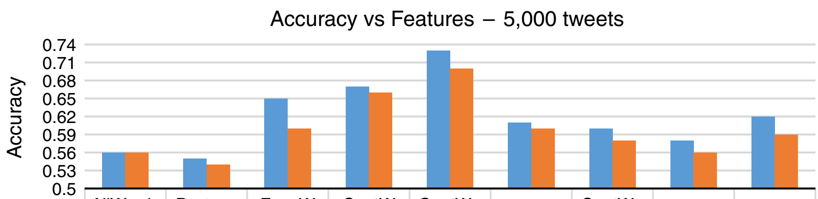


	AllWords	Pos tags	FuncW	ContW	ContW+FuncW	ngrampos	ContW+FW+ngram	ngrampos+FuncW	ngrampos+ContW
Accuracy-NV	0.58	0.58	0.59	0.59	0.63	0.58	0.6	0.58	0.61
Accuracy-ME	0.55	0.56	0.61	0.55	0.6	0.55	0.57	0.53	0.63

Features

Accuracy-NV Accuracy-ME

Figure 5.
Accuracy vs features – 3,000 tweets



	AllWords	Postags	FuncW	ContW	ContW+FuncW	ngrampos	ContW+FW+ngram	ngrampos+FuncW	ngrampos+ContW
Accuracy-NV	0.56	0.55	0.65	0.67	0.73	0.61	0.6	0.58	0.62
Accuracy-ME	0.56	0.54	0.6	0.66	0.7	0.6	0.58	0.56	0.59

Features

Accuracy-NV Accuracy-ME

Figure 6.
Accuracy vs features – 5,000 tweets

Table VI. Best differentiating features – part of speech *n*-grams

Part of speech– n-gram	Sarcasm/ non-sarcasm	Ratio	Meaning
WPS+BEZ	*Sarcasm*	*5.7:1.0*	*WH-pronoun, nominative + verb "to be," present, 3rd person singular*
NR$	*Non-sarcasm*	*5.5:1.0*	*Noun, singular, adverbial, genitive*
WPS	*Sarcasm*	*4.4:1.0*	*WH-pronoun, nominative*
UH	Non-sarcasm	3.7:1.0	Interjection
ABX	Non-sarcasm	3.3:1.0	Determiner/pronoun, double conjunction or pre-quantifier
HVG	Non-sarcasm	3.3:1.0	Verb "to have," present participle or gerund
EX	Sarcasm	2.2:1.0	Existential there
BEM	Non-sarcasm	2.1:1.0	Verb "to be," present tense, 1st person singular
BER	Sarcasm	2.0:1.0	Verb "to be," present tense, 2nd person singular or all persons plural

Table VII. Best differentiating features – function words

Function words	Sarcasm/non-sarcasm	Ratio
why	*Sarcasm*	*4.5:1.0*
very	*Sarcasm*	*4.5:1.0*
having	Non-sarcasm	3.8:1.0
got	Sarcasm	3.6:1.0
up	Sarcasm	3.4:1.0
ever	Non-sarcasm	3.2:1.0
until	Non-sarcasm	3.2:1.0
rather	Non-sarcasm	3.2:1.0
while	Non-sarcasm	3.2:1.0

non-sarcastic ones. Similarly, nouns which are singular, adverbial or genitive are five-and-a-half times more likely to be used by customers in non-sarcastic tweets than in sarcastic ones.

Table VII shows the function words that are the best differentiators between the sarcastic and the non-sarcastic sentences in our data set. For instance, the words "why" and "very" are four-and-a-half times more likely to be used in a sarcastic sentence by a customer as compared to a non-sarcastic one. It could be interpreted that the mentioned function words (in italics in Table VII) indicate that inquisitive and extreme tweets by customers are more likely to be sarcastic rather than non-sarcastic. We consider this to be an important finding of this study which could be of relevance to businesses which use online customer sentiments.

We summarize the major findings of our study as follows:

(1) Function words and content words together are the most important indicators of sarcasm in customer tweets in terms of accuracy and *F*-measure for both the classifiers.

(2) The Naïve Bayes classifier performs better than the maximum entropy classifier for most of the tweet data sets.

(3) Among the authorial style-based features function words are the best performers.

(4) Pronouns along with nominative verbs are more likely to be used in sarcastic tweets by customers than in non-sarcastic tweets.

(5) Extreme and inquisitive tweets by customers are more likely to be sarcastic.

The results obtained in Tables VI and VII are indicative and should not be deemed to be conclusive. The main focus of this study is to propose features which could be utilized in detecting sarcasm in online text. This study also claims that the proposed features are more effective than the features currently employed in detecting sarcasm.

The findings of this research work could be utilized by various organizations to label sentences for sentiment analysis, thereby improving the training set labeling correctness and subsequently improving the classification accuracy. Moreover, the training set size could be enhanced by including sentences which are sarcastic in nature. Also, the feature types introduced in this paper could be applied for sentiment analysis. This could potentially lead to interesting results, such as which part of speech is more likely to be used in positive sentences as compared to negative sentences.

5. Conclusion

Customers use twitter as a platform to express their emotions and opinions about social issues, products or services. Lexicon-based methods of text mining can at times fail to recognize sarcasm or provocative words used by customers online (Yee Liau and Pei Tan, 2014). For example, the word "thank you" can be used in a positive way, such as "Thank you Samsung for this wonderful phone #awesome" as well as in a negative way, such as "The phone battery doesn't even last an hour after charging to a 100%, Thank you Samsung." It is difficult to detect that the second sentence implicates a negative emotion using the current data mining tools. More often than not, these sentences end up being classified in the wrong category. Hence, there is a need of devising techniques which could detect sarcasm in tweets. This becomes even more crucial from a business perspective where online content is being increasingly used for customer management.

With the increasing complexity in the ways people communicate online, differentiating sarcastic text from non-sarcastic ones is crucial for correctly categorizing product/movie reviews (good or bad), social opinions (positive or negative) and even news (real or fake!!). Our research focuses on extracting features from tweets which can differentiate between a sarcastic and a non-sarcastic customer tweet by considering the author's writing style. Researchers in the past have attempted to detect sarcasm solely considering the content words used in the text. However, this could be misleading, and might lead to wrong interpretations or business decisions.

Hence, authorial or writing style-based features, such as function words and part of speech *n*-grams, which are largely ignored, are crucial for detecting sarcasm. This research adds a new dimension in the study of NLP as well as understanding online consumer behavior.

Our results reveal that including features that are independent of the text lead to an increase in the sarcasm detection accuracy. We also found that Naïve Bayes classifier performed relatively better than the maximum entropy classifier in differentiating sarcastic tweets from the non-sarcastic ones. Among the authorial style-based features, we observed that function words performed better than part of speech tags and part of speech *n*-grams.

This is a novel approach to sarcasm detection which we hope will lead to further studies using such features. In our study, we have considered a data set of 5,000 tweets. However, we still got an accuracy in excess of 70 percent, which is better than the accuracy levels achieved in extant literature. This clearly illustrates the relevance of the methods we have applied in this work.

The new approach to sarcasm detection could lead to an improvement in accuracy in case of supervised/unsupervised learning-based classification. This has direct implications on companies using text mining on online content to identify new customers, address customer problems and reduce customer churn rates.

In this work, we have used two different classifiers which are linear. This work could be extended to non-linear classifiers, such as SVM and perceptron and the results obtained could be compared to the ones obtained in our experimentation. One of the limitations of our current research is that we have not differentiated on types of sarcasm, instead we have considered it as a two-class problem. If the subtle variations in sarcasm could be identified and differentiated, it would lead to further refinement in categorizing customers and addressing customer problems more personally.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A.R. (2003), "Gender, genre, and writing style in formal written texts", *Text – Interdisciplinary Journal for the Study of Discourse*, Vol. 23 No. 3, pp. 321-346.

Argamon, S., Koppel, M., Pennebaker, J.W. and Schler, J. (2007), "Mining the blogosphere: age, gender and the varieties of self-expression", *First Monday*, Vol. 12 No. 9, available at: http://dx.doi.org/10.5210/fm.v12i9.2003

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2009), "Automatically profiling the author of an anonymous text", Vol. 52 No. 2, pp. 119-123, available at: http://doi.org/10.1145/1461928.1461959

Church, K.W. (1989), "A stochastic parts program and noun phrase parser for unrestricted text", *International Conference on Acoustics, Speech, and Signal Processing*, pp. 136-143.

Clark, H.H. and Gerrig, R.J. (1984), "On the pretense theory of irony", *Journal of Experimental Psychology*, Vol. 113 No. 1, pp. 121-126.

Davidov, D., Tsur, O. and Rappoport, A. (2010), "Semi-supervised recognition of sarcastic sentences in Twitter and Amazon", *Fourteenth Conference on Computational Natural Language Learning, July*, pp. 107-116.

Domingos, P. (2012), "A few useful things to know about machine learning", *Communications of the ACM*, Vol. 55 No. 10, p. 78.

Gergen, K.J. (1999), *An Invitation to Social Construction*, Sage.

Ghiassi, M., Skinner, J. and Zimbra, D. (2013), "Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network", *Expert Systems with Applications*, Vol. 40 No. 16, pp. 6266-6282.

Gibbs, R.W. (1986), "On the psycholinguistics of sarcasm", *Journal of Experimental Psychology*, Vol. 115 No. 1, p. 3.

Gibbs, R.W. and Colston, H.L. (2007), *Irony in Language and Thought: A Cognitive Science Reader*, Psychology Press.

Giora, R. (1995), "On irony and negation", *Discourse Processes*, Vol. 19 No. 4, pp. 239-264.

González-Ibánez, R., Muresan, S. and Wacholder, N. (2011), "Identifying sarcasm in Twitter: a closer look", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, Vol. 2 No. 1, pp. 581-586.

Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection", *The Journal of Machine Learning Research*, Vol. 3, pp. 1157-1182.

Ivanko, S.L. and Pexman, P.M. (2003), "Context incongruity and irony processing", *Discourse Processes*, Vol. 35 No. 3, pp. 241-279.

Jorgensen, J. (1996), "The functions of sarcastic irony in speech", *Journal of Pragmatics*, Vol. 26 No. 5, pp. 613-634.

Justo, R., Corcoran, T., Lukin, S.M., Walker, M. and Torres, M.I. (2014), "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web", *Knowledge-Based Systems*, Vol. 69, pp. 124-133.

Kacen, J.J., Hess, J.D. and Kevin Chiang, W.-Y. (2013), "Bricks or clicks? Consumer attitudes toward traditional stores and online stores", *Global Economics and Management Review*, Vol. 18 No. 1, pp. 12-21.

Katz, A.N., Blasko, D.G. and Kazmerski, V.A. (2004), "Saying what you don't mean: social influences on sarcastic language processing", *Current Directions in Psychological Science*, Vol. 13 No. 5, pp. 186-189.

Khan, A.Z., Atique, M. and Thakare, V. (2015), "Combining lexicon-based and learning-based methods for Twitter sentiment analysis", *International Journal of Electronics, Communication and Soft Computing Science and Engineering (IJECSCSE)*, pp. 89-91.

Klammer, T., Schulz, M. and Della Volpe, A. (2000), *Analyzing English Grammar, 6/e*, Pearson Education.

Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection", *International Joint Conference on Artificial Intelligence*.

Koppel, M. (2002), "Automatically categorizing written texts by author gender", *Literary and Linguistic Computing*, Vol. 17 No. 4, pp. 401-412.

Kreuz, R.J. and Glucksberg, S. (1989), "How to be sarcastic: the echoic reminder theory of verbal irony", *Journal of Experimental Psychology*, Vol. 118 No. 4, p. 374.

Kunneman, F.A., Liebrecht, C.C. and van den Bosch, A.P.J. (2014), "The (un) predictability of emotional hashtags in twitter", *Proceedings of the 5th Workshop of Language Analysis for Social Media*, pp. 26-34.

Manning, C.D. and Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge.

Martinez-Camara, E., Martin-Valdivia, M.T., Urena-Lopez, L.A. and Montejo-Raez, A.R. (2014), "Sentiment analysis in Twitter", *Natural Language Engineering*, Vol. 20 No. 1, pp. 1-28.

Maynard, D. and Greenwood, M.A. (2014), "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis", *LREC*, May, pp. 4238-4243.

Mukherjee, A. and Liu, B. (2010), "Improving gender classification of blog authors", *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, October*, pp. 207-217.

Mukherjee, S. and Bala, P.K. (2016), "Gender classification of microblog text based on authorial style", in Becker, J. and Shaw, M.J. (Eds), *Information Systems and E-Business Management*, Vol. 15, Springer, Berlin and Heidelberg, February, pp. 117-138, available at: http://doi.org/10.1007/s10257-016-0312-0

Ng, A.Y. and Jordan, M.I. (2002), "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes", *Advances in Neural Information Processing Systems*, Vol. 2, pp. 841-848.

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends® in Information Retrieval*, Vol. 2 No. 2, pp. 1-135.

Pennacchiotti, M. and Popescu, A.-M. (2011), "A machine learning approach to Twitter user classification", *ICWSM*, Vol. 11 No. 1, pp. 281-288.

Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M. (2010), "Classifying latent user attributes in twitter", *Proceedings of the 2nd International Workshop Search Mining User-Generated Contents*, p. 37.

Ravi, K. and Ravi, V. (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications", *Knowledge-Based Systems*, Vol. 89, pp. 14-46.

Rundell, M. and Fox, G. (2002), *Macmillan English Dictionary for Advanced Learners*, Macmillan, Oxford.

Saif, H., He, Y. and Alani, H. (2012), "Semantic sentiment analysis of Twitter", *The Open University's Repository of Research Publications Semantic Sentiment Analysis of Twitter Conference Item*.

Scherer, K.R. (2001), "Appraisal considered as a process of multilevel sequential checking", *Appraisal Processes in Emotion: Theory, Methods, Research*, Vol. 92, p. 57.

Schürer, S.C. and Muskal, S.M. (2013), "Kinome-wide activity modeling from diverse public high-quality data sets", *Journal of Chemical Information and Modeling*, Vol. 53 No. 1, pp. 27-38.

Thelwall, M., Buckley, K. and Paltoglou, G. (2011), "Sentiment in Twitter events", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 2, pp. 406-418.

Tsur, O. and Davidov, D. (2010), "Icwsm – a great catchy name: semi-supervised recognition of sarcastic sentences in product reviews", *Proceeding of the International (AAAI) Conference on Weblogs and Social Media*, pp. 162-169.

Turney, P.D. (2002), “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews”, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424.

Utsumi, A. (2002), “Toward a cognitive model of poetic effects in figurative language”, *2002 IEEE International Conference on Systems, Man and Cybernetics*, p. 6.

William Collins Sons & Co. Ltd (2009), *Dictionary-Complete, C.E. (1979). Unabridged 10th Edition 2009*©, William Collins Sons & Co. Ltd.

Winkler, E. (2012), *A Basic Course in Linguistics*, Bloomsbury Publishing.

Yan, X. and Yan, L. (2006), “Gender classification of weblog authors”, *AAAI Spring Symposium Series Oncomputational Approaches to Analysing Weblogs*, pp 228-230.

Yee Liau, B. and Pei Tan, P. (2014), “Gaining customer knowledge in low cost airlines through text mining”, *Industrial Management & Data Systems*, Vol. 114 No. 9, pp. 1344-1359.

Corresponding author
Shubhadeep Mukherjee can be contacted at: shubhadeep.mukherjee13fpm@iimranchi.ac.in