Knowledge-Based Systems 120 (2017) 15–33

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A novel automatic satire and irony detection using ensembled feature selection and data mining

Kumar Ravi[a,b], Vadlamani Ravi[a,*]

[a] Center of Excellence in Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad - 500057, India

[b] School of Computer & Information Sciences, University of Hyderabad, Hyderabad-500046, India

ARTICLE INFO

Article history:
Received 20 August 2016
Revised 5 November 2016
Accepted 19 December 2016
Available online 21 December 2016

Keywords:
Satiric news
Satire detection
Irony detection
Customer reviews
Ensembled feature subset selection
Sentiment analysis
LIWC
TAALES

ABSTRACT

Figurative language detection has always been a difficult task for human beings while being a more difficult proposition, even if automated using text and data mining. The available computational approaches are also quite limited in their capabilities and scope. In this regard, we propose an ensembled text feature selection method followed by a new framework in the paradigm of text and data mining to automatically detect satire, sarcasm, and irony found in news and customer reviews. The effectiveness of the proposed approach was demonstrated on three datasets including two satiric and one ironic dataset. The proposed methodology performed well on one satiric dataset and yielded promising results on the remaining two datasets. Moreover, we found out some interesting common characteristics of satire and irony like affective process (*negative emotion*), personal concern (*leisure*), biological process (*body* and *sexual*), perception (*see*), informal language (*swear*), social process (*male*), cognitive process (*certain*), and psycholinguistic (*concreteness* and *imageability*), which were extracted from three corpora. Of particular significance is the comparison of our approach with human annotators' evaluations, which served as a baseline in these tasks.

1. Introduction

Figurative language (FL) comprises irony, sarcasm, satire, parody, etc [1]. The study on the inter-relation between irony, satire, and parody can be found in [2]. The irony is often found in two forms like verbal and situational. In verbal irony, a speaker intends to express opposite meaning of what is being said. Situational irony implies to contradictory, juxtapositions, coincidences, counterfactuals, hypocrisy, etc. [1]. Satire and parody are two sub-types of irony and generally used for the same intention [3]. On the one hand, satire is mainly used to criticize an entity without explicitly mentioning its name or identity [1,4]. Here, an entity may refer to an institution, one or more individuals, social customs, actions, or beliefs [1]. On the other hand, parody is used to mimic or imitate an individual, who is being attacked with exaggerative act [5]. Moreover, another figurative device, called Sarcasm, is made up of positive words with negative intent and attacks target sharply. According to the Oxford dictionary[1] sarcasm is "The use of irony to mock or convey contempt."[2] The line between irony and sarcasm is very thin; hence linguists are often confused with these two terms.

FL is an integral part of any literal language. Due to the evolution of social media, the figurative language proliferated around us in various forms of online text such as forums [6], customer reviews [7], news [8], interviews [9], tweets [10], etc. Automatically, detecting FL is a quite cumbersome task, and hence, became a hot topic for linguistic and computational intelligence communities nowadays. Satiric and humorous content are often found in indirect speech, which should not be confused with a grammatical construct with the same name. An indirect speech contains sophisticated content, which implicitly conveys its message. Dealing with these kinds of contents found numerous applications in various areas such as psychology, philosophy, linguistics, cognitive science, sentiment analysis, etc. [1,10–12]. Some of the sentiment analysis tasks like review summarization [11,13], sentiment classification [14–17], etc. are severely affected by ironic and satiric content [3].

Ravi and Ravi [11] reported that figurative speech involving irony, sarcasm, and satire are inherent parts of customer reviews. These figurative devices often introduce vagueness in the polarity

* Corresponding author. Fax: +91-40-23535157.
E-mail addresses: ankitaravi.ravi00@gmail.com (K. Ravi), padmarav@gmail.com, rav_padma@yahoo.com (V. Ravi).

[1] http://www.oxforddictionaries.com/.

[2] http://www.oxforddictionaries.com/definition/english/sarcasm.

http://dx.doi.org/10.1016/j.knosys.2016.12.018
0950-7051/

of reviews, hence, severely affect the quality of the results of sentiment analysis. They also indicated that more rigorous experiments including appraisal theory based approaches could improve upon the obtained results. Journalist, Cartoonist, critics, etc. often use the satirical news to criticize and expose a political movement, an event, an individual, a product item, etc. The satirical news is a brilliant medium to point out flaws available in particular customs, rituals, policies, and regulations also. The satirical content often contains metaphor, analogy, ambiguity, and funny jokes, etc. [11,18]. A satirical text, a clever arrangement of words, can criticize attributes or behaviors of an entity with/without naming a particular object. However, the irony is used to deny what has been literally said [3].

In this study, we experimented with satiric news in the context of India and America as well as ironic reviews written on www.amazon.com. To see how satiric news looks like, let us consider some examples of the satiric news taken from my.fakingnews.firstpost.com.[3]

"...How to knot the tie should be a compulsory course as part of training curriculum and like fire drill, this should be repeated every quarter. Second, during campus placement and off campus recruitments, the interviewer should ask and check whether the job aspirant knows how to knot the tie? ... "

It was written in the context of the announcement of allowing informal dress in the office by a software company in India. This example attacks a newly formed rule in the software industry without mentioning the name of the enterprise. Let us consider one more example:

"Hisar. It has been more than 7 days since Haryana police is camping outside the Ashram of Rampal, the self-styled Godman. Unable to break the resistance of his supporters and enter the premise, Police has finally asked for help from self-styled expert in such matters, Inspector Daya. Rampal asking his supporters to push the door from inside. ACP Pradyuman confirmed the reports and assured that he will send inspector Daya to break open the door the way he always does. ..."

In this example, the satiric device is used to mock the "Haryana police". In order to target "Haryana police", the author suggested seeking help from the television serial characters like "Inspector Daya" and "ACP Pradyuman", which is literally making fun of the policing system of the given state. Let us see one more example from my.fakingnews.firstpost.com:

"...Holi: a Hindu festival of colors and secular festival of saving water."

In the first example, the first part of the sentence, "a Hindu festival of colors", is the literal language, whereas the second part, "secular festival of saving water." is used to contempt the existing customs by presenting the opposite of the reality. Since water is used in celebration of 'Holi'. Some satiric news from Burfoot and Baldwin [8] are presented below:

"...Winning is not everything in sport. How 'close' does a team come to winning or drawing a test match is equally important."

"Brazilian Judge awarded prize from Hitler Commemoration Institute"

"Syria attends mideast peace talks for free continental breakfast"

In these sentences, we can easily see the intention of the writer to ridicule on some sports' result, Brazilian Judge, and Syria. Some of the specific properties of irony are the *maxim of quality cooperative principle* [19], *echoic mention* [20], *pretence* [12], *the graded salience hypothesis* [21], *the fallacy of equivocation* [22], etc. The *maxim of quality cooperative principle* indicates that the speaker doesn't have sufficient evidence about what he said or what he speaks seems to be false. According to *echoic mention*, the speaker tacitly dissociates herself from an attributed utterance or thought. Under *pretence*, the irony speaker expects his audience to recognize the mocking or critical attitude behind him. The ironic speaker often crafts his utterances and highlights an expectation that has been violated by the target [23]. A few studies handled some of these features as described in Section 2. All these features together make irony and satire detection a difficult task.

Hence, we proposed a generic approach to automatically detect ironic content, and it's an important subtype satire [3]. After sufficient pre-processing of the raw content, we invoked from machine learning techniques. In this regard, we proposed a novel automated method for figurative content detection using an ensemble of unigrams, semantic, psycholinguistic, and statistical features and data mining techniques. The ensembled feature subset is fed as an input to a host of binary classifiers to determine satiric news and ironic customer reviews. Moreover, we extracted interesting features of satiric news and ironic reviews, which make them satiric or ironic. We also highlighted what kinds of features of satire and irony are common.

This remainder of the paper is organized as follows: Section 2 presents the literature survey of studies related to detection of figurative speech. Section 3 presents the proposed approach. The experimental setup is described in Section 4. Results and discussion are presented in Section 5. Human evaluation is presented in Section 6. The paper is concluded with some future research directions in Section 7.

2. Literature survey

The prominent research works carried out on the constituents of irony in the last two and half decades have been surveyed by Gibbs and Colston [1]. The nature, function, and understanding of irony are elaborated in Gibbs and Colston [1]. Some of the recent works on irony detection with respect to sentiment analysis are surveyed by Ravi and Ravi [11]. Moreover, Joshi et al. [24] reviewed some of the recent studies on sarcasm detection. Some of the related works in literature are reviewed here into two parts: (a) irony and satiric content detection related studies [2,3,7,8,12,21,25–28], and (b) sarcasm detection related studies [29–31].

To deal with figurative language, Burfoot and Baldwin [8] employed SVM to classify satirical news articles. They proposed a metric called VALidation (VAL), which relies on absurdity and unusual combination of named entities. They experimented with 4000 newswire documents and 233 satiric news articles. The serious news report was collected from English Gigaword Corpus, and the satiric news was crawled from Google manually. They reported the maximum F- measure of 79.8% for the given dataset. Reyes et al. [25] determined humor and irony in micro-blogs. They considered some linguistic and semantic properties of text like different kinds of ambiguity, polarity, unexpectedness, and emotional scenarios to detect them. They collected 10,000 micro-blogs from Twitter in each of the categories viz. humor, irony, politics,

[3] This link is subjective to change by the respective website administrator.

technology, and general. They performed binary classification of irony vs. others as well as humor vs. others and reported the highest F-measure of 93% for humor vs. irony and the lowest F-measure 56% for irony vs. general. Giora and Fein [26] considered different features of ironic texts like irony type ("familiar" and "less familiar"), context type ("ironically" and "literally biased"), stimulus type (word and non-word), and word type ("compatible" and "incompatible" with context). They employed 48 graduate students to comprehend ironic text and observed that familiar ironic text comprehension needed less time compared to less familiar one. They also found that salient meanings were more prominent than contextual information, which helps in faster comprehension of irony. Giora [21] reviewed the role of the graded salience hypothesis, the standard pragmatic model and the direct access view in the comprehension process of literal vs. figurative language. According to his review, there is no empirical support, which claims the literal and figurative language follow different processes for comprehension. Further, they concluded that the salience and the functionality are two features of the text, which can be used for deciding similarity and difference between literal and non-literal text. Hirsch and Blum–Kulka [9] employed the cooperative principle, echoic mention, and pragmatic insincerity for identifying irony in news interviews. Kreuz and Roberts [2] studied the relation between irony vs. satire, irony vs. parody, and satire vs. parody. They inferred that satire mainly consists of pretense, whereas, the echoic mention is an important part of the parody. Reyes and Rosso [3] invented three conceptual layers of irony namely signatures, emotional scenarios, and unexpectedness. These three layers, in turn, involve eight kinds of textual features to capture negation regarding irony devices. The signatures include pointedness in terms of punctuation marks and emoticons, counter-factuality involving contradiction and opposition denoting words like *about, nevertheless, nonetheless* or *yet* and their synonyms representing negation, and temporal compression like *suddenly, now* or *abruptly*, etc. Emotional scenarios include the degree of response like active and passive, the ease of preparing a mental picture of a word, and the level of pleasantness. The unexpectedness includes temporal and contextual imbalance. The temporal imbalance includes the divergences related to verbs only, whereas contextual imbalance defines inconsistencies within a context. They experimented with the polarity dataset v2.0 [32] of movie reviews, the polarity dataset v1.1 [33] of movie reviews, book reviews dataset [34], and satiric news dataset [8]. The degree of irony in the polarity dataset v2.0 was the highest. The polarity dataset v1.1 had the second highest degree of irony. The satiric news dataset has been placed on the third position. Finally, the book review dataset contains the least ironic content. They also found that positive movie reviews were more prone to irony than negative movie reviews. Using unsupervised learning, they reported that out of 10 web comments, 1 or 2 had figurative content. Out of 50 satiric news documents, they found out that 34 documents had satiric content. So, the true positive rate was 70.0% for the satiric news dataset.

Hao and Veale [12] proposed to classify ironic and non-ironic similes. They crawled 45,021 similies using Google API using search patterns like "as * as *", "about as * as *", etc. To determine the affective signature of each simile, they exploited Whissell's [35] dictionary of affect. They found that 71% of similes had positive sentiment oriented ironic effect, whereas 9% of similes had negative sentiment oriented irony affect. Similarly, the non-ironic similies contained the percentages of positive sentiment oriented and negative sentiment oriented similies as 9% and 12% respectively. Sulis et al. [27] established that there exists a difference between irony, sarcasm, and negation available in tweets. They experimented with the dataset crawled for the task 11 of a SemEval-2015 challenge. The distribution of tweets is #not (3247 tweets), #sarcasm (2260) and #irony (1737). They exploited 14 dictionaries to observe the distribution of different semantic and psycholinguistic information available regarding *irony, sarcasm*, and *not*. By that, they inferred that irony was more creative and implicit than sarcasm. Sarcasm contains more positive sentiment than irony. Negation had been used for self-mockery and different from sarcasm and irony. To express negative intention, sarcasm is very often used along with positive words, whereas the converse is not true. They performed binary classification for irony vs. sarcasm, irony vs. not, and sarcasm vs. not, and yielded the F-measure of 69.8%, 75.2%, and 68.4% respectively using Random Forest (RF) [36]. They considered 10,000 tweets for each of the hashtags #not, #irony, and #sarcasm. Charalampakis et al. [28] employed semi-supervised and supervised algorithm to detect irony in 61,427 tweets posted on Greek elections. They considered five kinds of features namely spoken, rarity, meanings, lexical, and emoticons. They employed a host of classifiers for supervised classification and collective-tree for semi-supervised training. The functional trees yielded the best precision of 82.4% for supervised classification whereas RF yielded the best precision of 83.1% for semi-supervised validation.

Tsur et al. [29] proposed a semi-supervised approach to automatically classify sarcastic and non-sarcastic sentences. In their proposal, they developed a list of sentence patterns based on high- and low-frequency words to determine sarcastic sentences. In addition to that, they considered syntactic features like the length of a sentence, frequency of special and different case letters, etc. They experimented with 66,000 reviews for 120 products taken from www.amazon.com. Justo et al. [30] proposed to classify sarcasm and nastiness available in dialog texts. They experimented with sarcastic and nasty labeled parts of Internet Argument Corpus (IAC) [6]. They proposed a list of various combination of features like a list of Mechanical Turk cues, statistical cues, unigrams, bigrams, trigrams, the distribution of n-grams of Parts-of-Speech (POS), semantic features using Linguistic Inquiry Word Count (LIWC) [37] and sentiment-oriented features using SentiWordNet 3.0 [38], length information, the polarity of concept retrieved from SenticNet 3 [39]. They employed χ^2 statistics for features subset selection. The binary classification was performed using rule-based and Naïve Bayes (NB) classifier. NB yielded the best recall of 77.4% using two combinations of features for sarcasm detection. Both combinations of features yielded the same recall. The first combination of features was statistical cues and length information. The second combination of features was statistical cues and semantic features of LIWC, which yielded the best recall of 84.7% for nastiness detection also. Kunneman et al. [31] performed sarcastic tweets classification. They employed Balanced Winnow [40] classifier to classify 406,439 sarcastic and 406,439 non-sarcastic tweets. They tested the model on a small dataset of 353 tweets and reported an Area Under Curve (AUC) of 85.0%.

In above-mentioned works, we figured out six kinds of limitations: (a) the performance of the proposed models is not so good in [3,7,8,25,27], (b) the number of features was too large [27], (c) searching each word in a sentiment dictionary was a cumbersome task [27], (d) dependency on annotators for features extraction [26,27], (e) the test dataset was too small [31], and (f) extraction of POS features [36]. If we search each word in a sentiment dictionary, or we perform POS tagging, it will be a quite time-consuming approach. Similarly, relying on annotators for feature extraction is a cumbersome approach and introduces subjectivity.

Hence, we developed a new approach to address some of these issues. If we consider the rest of the literature, they are not directly comparable to our study due to two reasons i.e. either the experimental domain was different [25,27,28,31] or different kinds of figurative devices were handled [12,29–31]. Despite that, we included them into the literature survey, because the approaches needed to detect them share similarities with irony or satire detection

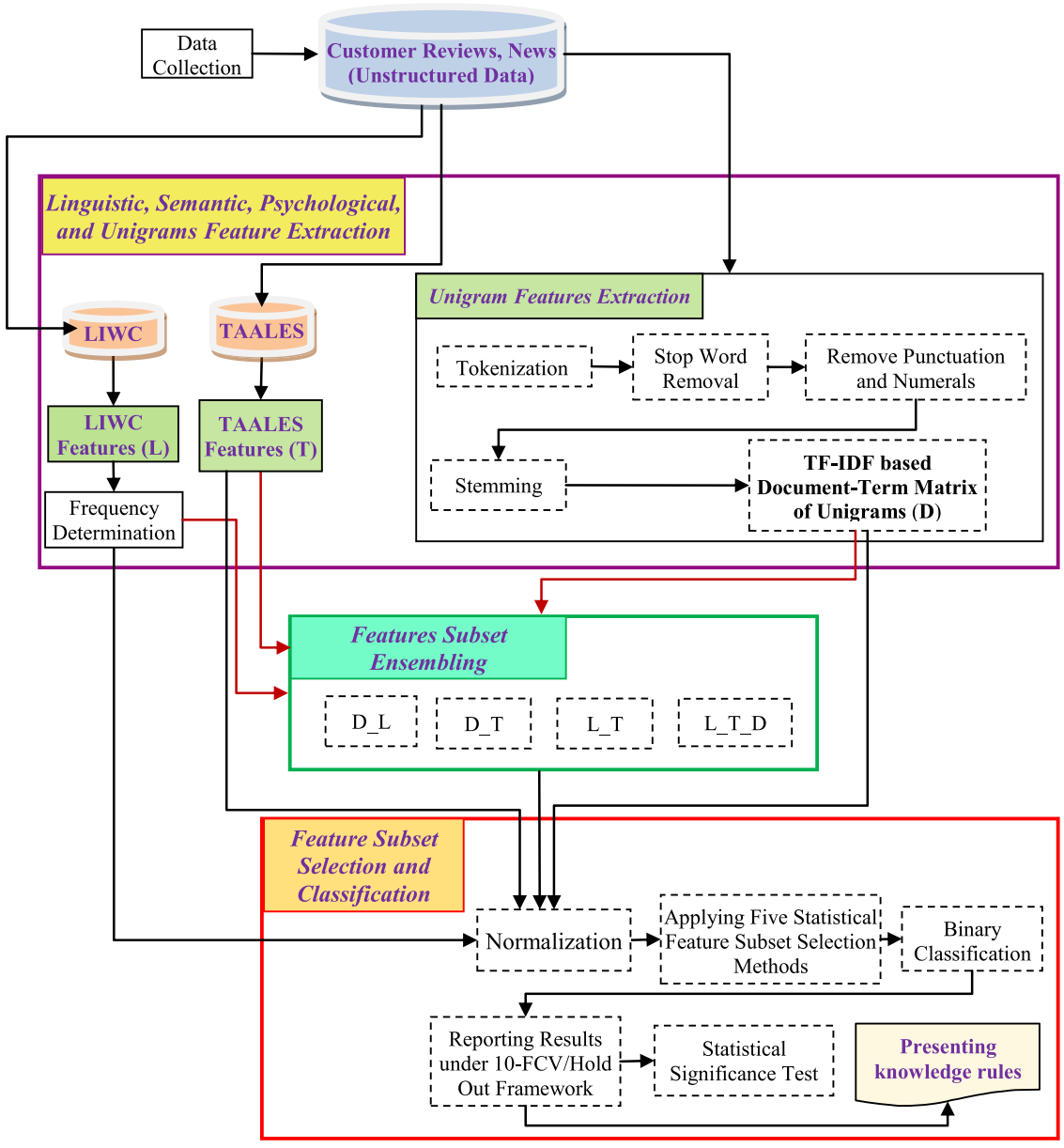

Fig. 1. The proposed approach.

approaches. Further, various kinds of figurative devices share some common characteristics.

3. Proposed approach

The schematic of the proposed approach is depicted in Fig. 1. The proposed approach comprises four phases: (i) data collection, (ii) linguistic, semantic, psychological, and unigrams feature generation, (iii) feature subset ensembling, and (iv) features subset selection and classification. The details of data collection are presented in Section 4.1, and the rest of the above are described in this section. In order to identify ironic customer reviews and satiric news, we developed an ensemble of semantic, psycholinguistic, and statistical unigrams feature, which is then used to develop a binary classification framework. The binary classification based framework comprises normalization, followed by feature selection with the help of five methods namely Information Gain (IG), Gain Ratio (GR), χ^2-statistics (CHI), correlation (CORR), and t-statistic (TSTAT), and a host of classifiers involving SVM using four kernels: (i) Linear, (ii) RBF, (iii) Sigmoid, and (iv) Polynomial [41,42], LMT [43], Logistic Regression [44], RF, NB [45], Bayesian Network, Multilayer Perceptron (MLP) [46], C4.5 [47], Classification and Regression Trees (CART) [48], and RIPPER [49].

3.1. Linguistic, semantic, psychological, and unigrams feature extraction

In order to capture linguistic, semantic, and psychological features from the news and reviews, we employed two dictionaries: Linguistic Inquiry and Word Count-2015 [37,50] and Tool for the Automatic Analysis of LExical Sophistication (TAALES) [51] which are described in Section 3.1.1. We obtained unigrams features using data mining techniques as presented in Section 3.1.2.

3.1.1. Linguistic, semantic, and psychological features

In addition to *psychological* features, LIWC also generates the distribution of *summary language, grammatical*, and *linguistic* features. LIWC tool uses LIWC dictionary at the backend containing

6400 words, which receives a corpus of text as inputs and generates a count of words into pre-specified psychologically meaningful categories as output. It is noticed that a word is counted into multiple categories. LIWC provides the psychometrics of words usage, content, and style words, etc. The content words included nouns, regular verbs, adjectives, and adverbs. Style or function words involve articles, auxiliary verbs, conjunctions, pronouns, prepositions, etc. From a psychological viewpoint, function words convey the style of communication, whereas content words indicate the expressed fact. Hence, we can say that style words demonstrate measures of people's social and psychological worlds. Overall LIWC provides three categories of words i.e. summary language (analytical thinking, clout, authenticity, and emotional tone), linguistic dimensions (function words, negation, etc.), and psychological processes (affective processes, social processes, cognitive processes, perceptual processes, biological processes, drives, time, relativity, personal concerns, and informal language) [50]. LIWC 2015 also takes into account abbreviated text like "b4" and emoticons like ":)". The captured features of LIWC 2015 can distinguish between different contexts tool [50]. In total, LIWC 2015 produces 95 features, which are considered as one set of features for our proposed method. Henceforth, these features will be referred as *L*.

Although LIWC 2015 considers 41-word categories such as tapping psychological constructs, it doesn't capture reference corpus frequency, range counts, or psycholinguistic norms such as *concreteness* or *imageability*. Therefore, TAALES [51] is mainly developed to measure lexical sophistication of a language. To accomplish that TAALES generates various features from pre-specified corpora for the input dataset. The features include corpus reference frequency, range, n-gram frequencies, academic lists, and psycholinguistic features. The frequency of a word with respect to a reference corpus appeared as a whole or in the most frequent lists like 1000 words, 2000 words, etc. The range is the count of documents of a corpus, which contains a particular word. The n-gram frequencies refer to the number of bi-grams and tri-grams with reference to some standard corpus like British National Corpus (BNC). The academic lists contain frequent words appearing in academic literature like journals, textbooks, etc. but not in spoken languages. The psycholinguistic properties include *concreteness, familiarity, imageability, meaningfulness*, and *age of acquisition*. The concreteness depends upon the degree of simplicity to describe a word like *a parrot* vs. *infinity*. The familiarity refers to the judgment of popularity of a word among adults and correlation with frequency counts. The *imageability* measures the degree of complexity of imagination about a word like *bridge* vs. *egress*. The ease of relating a word to another word is called *meaningfulness*. The *age of acquisition* indices are based on judgments about at what particular age a person learns a word. In total, TAALES generates a set of 105 feature indices for any corpus [51], which is considered as the second set of features to capture characteristics of satire and irony in the present study. The features obtained using TAALES will be referred as *T* hereafter.

3.1.2. Unigrams features extraction

We considered a set of unigram features under this category to obtain a TF-IDF representation based Document-Term Matrix (DTM) since it yields promising results in the area of sentiment analysis [33]. In the case of sentiment classification, and satire and irony detection, DTM provides content based details of a corpus and is used as a summarization of a corpus too. To generate n-gram features, we need text preprocessing steps like tokenization, stop word removal, remove punctuation and numerals, stemming, and DTM generation. Text preprocessing performs these tasks to convert unstructured data into structured data. Among them, the tokenization is a process to break a sentence into words, phrases, symbols or other meaningful tokens by removing punctuation marks. Stop words includes preposition, pronoun, and some other words, which occur so often and do not contribute to the analysis. Therefore, these terms were dropped. The punctuation marks and numerals are not considered under this category of features. Stemming improves the effectiveness of the method by bringing each word into its root form by truncating suffixes. The filtered text is converted into DTM, where each document vector comprises TF-IDF value as its elements. We considered TF-IDF value since it captures the relative importance of each term with respect to the whole corpora. The DTM is considered as the third feature set to capture additional content based characteristics of the satire and irony in a given text. From now on, the DTM features will be referred as *D*.

3.2. Feature subset ensembling

In addition to stylometrics, semantic metrics, and psychometrics, content-based metrics like DTM play a significant role in linguistic and sentiment analysis of the text. The stylometrics includes the distribution of capital letters, small letters, POSs, punctuations, numerals, the length of words, phrases and paragraphs, etc. The semantic metrics include the distribution of sentiment polarity words like the positive and negative sentiment, emoticons, and five primary emotions like love, joy, surprise, anger, and fear proposed by Parrott [52]. In order to capture different dimensions of characteristics of a text corpus, we ensembled all these sets of features together. We conjecture that the ensembled feature subsets obtained in various combinations improve final prediction results. Furthermore, the ensembled features provide useful insights of satiric and ironic content, which is independent of all corpora. Accordingly, we considered all possible combinations of *D, T*, and *L* such as *union of D and L (D_L), union of D and T (D_T), union of L and T (L_T),* and *union of L, T, and D (L_T_D).*

3.3. Feature subset selection and classification

Before applying an Feature Subset Selection (FSS), we first apply normalization to scale down the values of the seven sets of input features. The max-min normalization was applied to each feature set to bring them in the range of [0, 1].

3.3.1. Feature subset selection

We applied five FSS methods on each of the ensembled seven feature sets to identify and remove non-discriminating features. The employed FSS methods are gain ratio [47], information gain [53,54], t-statistic [55,56], χ^2-statistics [57], and correlation based [58]. The TSTAT method requires computation of t-statistic for every feature. The formula for computing t-statistic presented in Eq. (1).

$$t - statistic = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

where μ_1 and μ_2 are mean values of a feature column for class 1 and class 2 respectively, $\sigma_1$and σ_2 are standard deviations for two classes, n_1 and n_2 are the number of samples available in two classes. The *t-statistic* will be calculated for each feature of the feature set and a feature having the highest *t*-value will be considered at first and so on.

3.3.2. Employed classification methods

We employed a host of classifiers like LibSVM using four kernels namely RBF (SVMR), Sigmoid (SVMS), Linear (SVML), and Polynomial (SVMP) [42], LMT [43], Logistic Regression (LR) [44], RF, NB [45], Bayesian Network (BN), Multilayer Perceptron (MLP) [46],

C4.5 (J48), Classification and Regression Trees (CART) [48], and RIPPER (JRip) [49]. In order to gain insights into satire and irony in a text, we invoke transparent machine learning algorithms such as C4.5, CART, and LMT, which yields classification rules. Therefore, this section presents a brief introduction to these algorithms to make this paper self-contained.

3.3.2.1. Support vector machine. SVM was proposed by Cortes and Vapnik [41]. It is suitable for the high dimensional dataset. SVM works on the principle of maximum-margin classifier, which fits hyperplanes among samples by projecting them to higher dimensions. It yields a list of support vectors, which is a subset of samples lying on the decision boundaries. The SVM without any kernel function is known as linear SVM, which is the most common form. SVM is based on solving following quadratic optimization problem.

$$\min_{\omega,b,\xi} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i$$

subject to $y_i(w^T\phi(x_i)+b) \geq 1-\xi_i$, and $\xi_i \geq 0, i = 1, ..., l$,

where $\phi(x_i)$ maps x_i into a higher-dimensional space and $C > 0$ is the regularization parameter. The $\phi(x_i)^T\phi(x_j)$ is the kernel function can be represented using $K(x_i, x_j)$. The most commonly used kernels are:

(i) Linear: $K(x_i, x_j) = x_i^T x_j$,
(ii) Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$,
(iii) Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j\parallel^2), \gamma > 0$, and
(iv) Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$. Here, γ, r, and d are kernel parameters.

3.3.2.2. Logistic regression. Logistic regression [59] is developed analogous to linear regression. The linear regression fits a straight line on a set of samples using continuous target variable, whereas LR fits an *S*-shaped curve using discrete target variable. LR yields the distribution of probabilities that a sample will belong to an m number of classes. The ridge based LR was proposed by Cessie and Houwelingen [44] in 1992. The ridge is used for regularization purpose to save from overfitting of the model. This technique is too popular to be described in great detail here.

3.3.2.3. Logistic model tree. Logistic Model Tree (LMT) has been developed by combining the working principle of the decision tree and logistic regression [43]. Logistic regression is employed at every leaf node of a decision tree. LMT starts to create a decision tree by fitting a simple linear regression function to root node, which will be repeated for a determined number of times using the LogitBoost algorithm. The number of repetition of fitting a linear regression is decided using 5-Fold Cross Validation (FCV). For each node, either binary or multi-way splits can be employed. And these processes are repeated for all children nodes in an iterative manner until some stopping criteria are met. At the completion of the tree, pruning is performed using classification and regression tree algorithm. In order to classify a test sample, LMT uses Eq. (2) to determine the probability of different classes.

$$\Pr(G = j \mid X = x) = \frac{e^{F_j(x)}}{\sum_{k=1}^{J} e^{F_k(x)}} \tag{2}$$

where G represents the number of classes of the dataset, X accounts for a sample set, and

$$F_j(x) = \alpha_0^j + \sum_{v\epsilon V_t}\alpha_v^j . v.$$

In Eq. (2), α_v^j represents a coefficient of a feature or variable, v, for j^{th} class and V_t refers to the number of features available in the dataset. Compared to other decision trees, it cannot handle missing data, so before applying the model, we should impute mean or mode at the place of missing values. In terms of computational complexity, it involves a bit more complexity than a logistic regression.

3.3.2.4. C4.5. C4.5 is a decision tree proposed by Quinlan [47]. It is a non-parametric approach to machine learning. It constructs a tree in top-down and recursive divide-and-conquer manner using greedy approach. It uses information gain of attributes to determine the splitting criterion. The best information gain value of an attribute is selected for the root node and its number of branches is equal to the distinct values of that attribute. Each leaf node represents a class label. A classification rule is formed by using values lying on the path from root node to a leaf node. A classification rule provides the range of values of a feature for classifying a sample.

3.3.2.5. Classification and regression tree. Classification and Regression Trees (CART) was proposed by Breiman et al. [48]. It also follows the greedy, top-down, and divide-and-conquer approach to building the binary decision tree. This tree can be used for classification as well as regression purpose. The Gini index is used as the splitting criterion to identify an attribute to be considered for a node. The Gini index measures impurity exist in the data using Eq. (3)

$$Gini(D) = 1 - \sum_{j=1}^{m} p_j^2, \tag{3}$$

where p_j is the probability that a sample belongs to the class C_j. The p_j is computed using $|C_{i,D}|/|D|$, i.e. the ratio between the number of samples of class I and the total number of samples of the dataset

3.3.2.6. Repeated incremental pruning to produce error reduction (RIPPER). RIPPER generates rules using incremental reduced-error pruning [49]. The stopping criteria depend on the description length of the examples and rule set. The description length depends on the number of bits needed to send a set of examples with respect to a set of rules. The induction of new rules should ensure the reduction of length of the rules. It performs well on large datasets.

3.3.2.7. Random forest. Random forest was proposed by Breiman by ensembling decision trees [36]. It fits multiple decision trees on data by selecting N samples randomly, with replacement, from the training dataset. In order to fit multiple trees, RF randomly selects F input features to split at each node of the decision tree. The value F is commonly set as $\log_2(|D|) + 1$, where $|D|$ is the total number of samples. The small value of F will yield less correlation among decision trees whereas the bigger value of F will produce the better classification accuracy. The results for multiple decision trees are combined using Bagging.

3.3.2.8. Naïve Bayes. NB [45] is based on Baye's rule and is best suitable for the moderate or large size of the training dataset. It assumes strong conditional independence among features, given the target label. Conditional independence among features allows the classifier to estimate the class-conditional probability of only given a sample, instead of every combination of all samples of the dataset. This technique is too popular to be described in great detail here.

Table 1
Burfoot and Baldwin [8] dataset details.

	#Training samples	#Test samples	Total
True news	2505	1495	4000
Satiric news	133	100	233

3.3.2.9. Bayesian network. Bayesian network is a Directed Acyclic Graph (DAG), where each node of the graph represents a random variable in the form of conditional probability table (CPT). Here, each feature of the dataset is a random variable. Initially, it defines two algorithms to develop the learning algorithm; first, a function to determine the network for given data, second, a function to search through the space of possible networks. The log-likelihood of the BN is estimated by multiplying probabilities of all instances, where the probability is assigned for each instance on the network. An entry of a CPT is determined using relative frequencies of the associated combinations of attribute values in the training data.

3.3.2.10. Multilayer perceptron. MLP is proposed by Rumelhart et al. [46]. It used backpropagation algorithm to learn the weights of the neural network. It found various applications for classification, regression, and function approximations like spam detection, FOREX rate prediction, sentiment classification, etc. This technique is too popular to be described in great detail here.

4. Experimental setup

The experiments were carried out on three datasets. The first two datasets are related to the satiric news, and the third dataset is related to the ironic customer reviews. This section presents the details of datasets, feature subsets, classification and fine-tuning model parameters, and model validation method.

4.1. Data collection

We experimented with two satiric datasets and one ironic dataset. The first satiric dataset was obtained from Burfoot and Baldwin [8], and its distribution is presented in Table 1. Subsequently, Burfoot and Baldwin [8] dataset will be referred as BUR dataset. We developed the second satiric dataset on the satiric news in the Indian context and third dataset on ironic reviews crawled from www.amazon.com. For the second dataset, we developed a web crawler to crawl and collect serious news from www.firstpost.com and satiric news from my.fakingnews.firstpost.com. We collected 1272 serious and 393 faking news from respective websites, from now on it will be attributed as FAKE dataset. The faking news was written on different areas like Indian politics, sports, events, movies, etc. For our classification frameworks, we treated serious news as negative samples and faking news as positive samples. For irony detection, we employed our proposed model on customer reviews collected from www.amazon.com. We collected 2498 serious reviews and 499 ironic reviews. Hereunto, this dataset will be known as AMA. Here, serious reviews were considered under the negative class and ironic reviews under the positive class. The distribution of customer reviews on various products is presented in Table 2. Ten products listed in Table 2 were considered from Skalicky and Crossley [7]. Some of the products listed under ironic reviews category have also been considered as ironic in [17,60]. Reyes and Rosso [17] considered all five stars rating based reviews as ironic. Payton and Weigandt [60] reported that all reviews on "Hutzler 571 Banana Slicer", "BIC Cristal For Her Ball Pen", "Images SI Inc Uranium Ore", etc. are ironic. Instead of considering all reviews as ironic, we selected ironic reviews in two steps. We collected 500 reviews on each of five products with ironic reviews at the beginning. At the next step, we filtered ironic reviews for each of the products. For serious reviews, we collected 1000 reviews on each product and then we selected 500 non-ironic reviews on each of the products. All these datasets can be obtained by sending an e-mail to the corresponding author.

Table 2
Ironic news details.

S#	Serious reviews	#Reviews	Ironic reviews	#Reviews
1.	Carhartt Workwear Short Sleeve Original K87	500	Mountain Three Short Sleeve T-Shirt	99
2.	Paderno World Cuisine A4982799 Tri-Blade	500	Hutzler 571 Banana Slicer	100
3.	Paper Mate Retractable Ballpoint 1,904,804	500	BIC Cristal For Her Ball Pen	100
4.	Switch Sparkling Juice Variety 8 Ounce	498	Tuscan Dairy Whole Vitamin Gallon	100
5.	Weber 7416 Rapidfire Chimney Starter	500	Images SI Inc Uranium Ore	100
	Total	2498		499

Table 3
The feature sets and their size for different datasets.

S#	Dataset name	Feature sets and their size						
		D	*L*	*T*	*D_L*	*D_T*	*L_T*	*D_L_T*
1.	BUR	48,946	93	103	49,039	49,049	196	49,142
2.	FAKE	20,518	93	103	20,611	20,621	196	20,621
3.	AMA	10,540	93	103	10,633	10,643	196	10,643

4.2. Feature subset ensembling and feature subset selection

LIWC (L) generated 95 features, out of them *filename* and *segment* have been dropped for our experiments. The filtered 93 features of LIWC were considered as the first feature set for the experiments. The class label has been assigned as an additional feature. TAALES (T) produced 105 features; out of them, we considered 103 features after dropping two features i.e. *AWL.Sublist.10.Normed* and *filename*. *T* is considered as the second feature set for the whole experiments; we appended the class label as an extra feature. To obtain unigrams features, we employed text mining package "tm" of R to perform required text preprocessing on three corpora [61]. We employed Snowball stemmer for stemming [62,63]. The number of features obtained using DTM is referred as *D*. So, we obtained three feature sets namely *D, T*, and *L* as presented in Section 3.1.1 and 3.1.2. Now, these three feature sets were combined to obtain *D_L, D_T, L_T*, and *L_T_D*. Finally, we obtained in total seven set of features for each corpus resulting in seven datasets. The seven sets of features is arranged alphabetically as (a) *D*, (b) *D_L*, (c) *D_T*, (d) *L*, (e) *L_T*, (f) *L_T_D* and (g) *T*. The distribution of feature subsets is presented in Table 3. On each of the datasets so obtained, we performed feature selection using IG, GR, CORR, CHI, and TSTAT. The datasets obtained from the reduced features were then classified using ten classifiers as listed in section 3.6.

4.3. Classification and fine-tuning model parameters

Ten classifiers listed in section 3.6 are available in Weka 3.7.13 [64], which we used for binary classification. We used *Weka Experimenter* to perform 10-FCV experiments using default parameter settings of each classifier for FAKE and AMA dataset, whereas experiments were carried out under hold-out framework for BUR

Table 4
The range of parameter values of each model used in grid search.

S#	Model name	Parameter abbreviation	Parameter name and/or purpose	Default value	Range of parameter values	Step size
1.	**LMT**	-I	numBoostingIterations	−1	[−1, 8]	1
		-M	minNumInstances	15	[5, 25]	5
		-W	weightTrimBeta	0.0	[0.0, 0.2]	0.05
2.	**LR**	-R	Ridge	1.3E-8	[1.0E-12, 1.0E2]	1
		-M	maxIts: maximum number of iterations to perform	−1	10	–
3.	**J48**	-C	confidenceFactor	0.25	[0.15, 0.35]	0.05
		-M	minNumObj	2	[2, 5]	1
4.	**CART**	-M	minNumObj	2	[2, 5]	1
5.	**JRip**	-N	minNo: The minimum total weight of the instances in a rule	2	[1, 5]	1
6.	**RF**	-K	numFeatures: number of features considered for building a decision tree of the random forest	log(numAtt) + 1	If numAtt <= 32 then [log(numAtt) + 1, log(numAtt) + 5] If N > 32 then [log(N) + 1, log(N) + 11]	1 1
		-I	numTrees	100	[70, 130]	10
7.	**MLP**	-L	learningRate	0.3	[0.02, 0.25] and 0.3	0.04
		-M	Momentum	0.2	[0.2, 0.9]	0.2
		-N	trainingTime	500	[500, 1250]	250
		-H	Number of hidden layers (NHL), number of nodes (NHN) per hidden layer	NHL: 1, NHN: 10	NHL: 1, if numAtt <= 11 then NHN: [5, 8] NHL: 1, if numAtt <= 31 then NHN: [10, 20] NHL: 1, if numAtt > 31 then NHN: [numAtt/2 - 10, numAtt/2 + 10]	3 5 5
8.	**SVML**	-C	Cost	1.0	2^c where c = −3 to 15	2
9.	**SVMP**	-C	Cost	1.0	2^c where c = −3 to 15	2
		-D	Degree	3.0	[3, 9]	1
		-G	Gamma	0.0	2^g where g = −15 to 3	2
		-R	Coefficient	0.0	2^r where r = −5 to 10	2
10.	**SVMR**	-C	Cost	1.0	2^c where c = −3 to 15	2
		-G	Gamma	0.0	2^g where g = −15 to 3	2
11.	**SVMS**	-C	Cost	1.0	2^c where c = −3 to 15	2
		-G	Gamma	0.0	2^g where g = −15 to 3	2
		-R	Coefficient	0.0	2^r where r = −5 to 10	2

dataset. Each classifier with default parameters was applied on each of the feature subsets in Section 4.2. Out of those feature subsets, we selected the best number of features for each feature selection method to further tune the parameters of each classifier. Using the best number of features for each dataset, the parameters of the ten classifiers were tuned using *Weka API* and Java programming. So, we tuned the parameters of the ten classifiers to improve upon the obtained average of 10-FCV AUC. We employed grid search technique to determine the best combinations of parameters for each classifier. Table 4 presents the list of models, their parameter names, and the range of parameter values experimented with. The *parameter abbreviation* column represents the name of the alphabet used by Weka 3.7.13 for different parameters. It should be noted that only these parameters of the given algorithm were fine tuned and not every other parameter. The name and/or purpose of each parameter are presented in the next column. The *default value* column presents the default value of each parameter pre-specified in Weka. The default value of each parameter was considered as a reference point to determine the range of values for each parameter as listed in the next column. Finally, the last column *step size* indicates an amount of increment to each parameter value, each of which was used to perform experiments.

4.4. Model validation

In order to measure the performances of different classifiers and feature sets, we used sensitivity/recall (SEN), specificity (SPE), Area Under Curve (AUC), precision (PREC), and F_1. We presented an average SEN, average SPE, average AUC, average PREC, and average F_1 of 10-Fold Cross-Validation (FCV) for a classifier for FAKE and AMA datasets. In the case of BUR dataset, we reported these performance measures using the hold-out method. For the hold-out procedure, the distribution of training and testing parts is presented in Table 1. To detect irony and satire, we considered satiric news or ironic reviews under positive class, whereas serious news or reviews were considered into negative class. For a task like ironic and satiric content detection, the positive class is crucial for a classifier to predict successfully. Since the amount of satiric and ironic content compared to serious news and reviews will be too minimal on the web. Therefore, detection of the positive class will be quite important and a difficult task. Hence, in our case, SEN is more important than SPE. In order to have a holistic picture of the overall performance of a classifier or a feature set, we reported AUC also and we compared different feature sets and classifiers in terms of AUC.

5. Results and discussion

To report the performance of different classifiers, we indicated the best parameter values wherever applied; otherwise, the default parameter value was used. Table 5 presents the best AUC of the hold-out method for BUR dataset for each feature set. Table 6 presents the performance of the proposed method by Burfoot and Baldwin [8] on BUR dataset. They employed two kinds of feature representation for unigram based DTM namely the binary feature weights (BIN) and bi-normal feature scaling (BNS). They applied both feature representations on two sets of features and their combinations viz. lexical (LEX), VAL, and LEX+VAL (ALL). The best PREC and F_1 reported by them were 95.8% and 79.8% respectively, whereas we obtained the best PREC and F_1 of 96.97% and 85.86% respectively. LR using TSTAT based 500 features of *L_T_D* yielded the best AUC and F_1 of 92.07% and 85.86% respectively as presented in Table 5. Table 7 presents a comparison among different model using maximum 300 features selected using different feature selection method. LR outperformed the remaining classifiers and yielded the best AUC of 88%.

Table 5
Results of various feature combinations for BUR dataset.

FSet[a]	Model	FSS	NF[b]	Model parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)
D	NB	IG	80	Default	62.0	93.31	77.66	38.27	47.33
D_L	LR	TSTAT	300	-R 0.1 -M −1	85.0	97.92	91.46	73.28	78.70
D_T	LR	GR	300	-R 1.0E-5	64.0	99.86	81.93	**96.97**	77.11
L	SVML	ALL	93	-C 524,288.0	77.0	98.52	87.76	77.78	77.39
L_T	LR	CORR	120	-R 0.001	78.0	97.72	87.86	69.64	73.59
L_T_D	LR	TSTAT	500	-R 0.1 -M 10	85	99.13	**92.07**	86.73	**85.86**
T	SVMP	CHI	10	-D 9 -G 8.0 -R 0.03125 -C 0.5	73.0	83.34	78.17	22.67	34.6

a. Feature Set, b. Number of features

Table 6
The results reported by Burfoot and Baldwin [8] for BUR dataset.

("article → SATIRE?")	PREC (%)	Recall (%)	F_1 (%)
All-positive baseline	6.3	**100**	11.8
BIN	94.3	50.0	65.4
BIN + LEX	94.5	52.0	67.1
BIN + VAL	94.3	50.0	65.4
BIN + ALL	94.5	52.0	67.1
BNS	94.4	67.0	78.4
BNS + LEX	95.7	66.0	78.1
BNS + VAL	94.5	69.0	**79.8**
BNS + ALL	**95.8**	68.0	79.5

Tables 8 through 12 present results obtained for FAKE dataset using seven feature set combinations and thirteen classifiers under the 10-FCV framework. Table 8 presents the results of various feature combinations. The LMT yielded the best AUC of 97.43% using 30 CHI based features of L with given parameter values. We performed *t*-test among seven feature sets over AUC. Throughout the paper, *t*-test is performed at 1% level of significance and 18 degrees of freedom. The computed t-statistic (t-stat) values, which are less than 2.83, indicate that the AUC obtained using D_L, L, L_T and L_T_D are statistically significantly same. So, it indicates that LIWC is the most important contributor in determining satire for FAKE dataset. For each of the best feature sets, we compared the performance of ten classifiers. Table 9 presents the comparison of results using D_L feature set. LMT yielded the best AUC of 97.3% on 200 CHI based features and default parameter values. With the help of *t*-test, we found that J48 using CHI based 20 features, JRip using CORR-based 100 features, LMT using CHI based 200 features, LR using TSTAT-based 40 features, MLP using TSTAT based 20 features, RF using TSTAT based 15 features, and CART using CORR based 160 features yielded statistically the same AUC. Table 10 presents the comparison of results using *L* feature set. For the *L* feature set, all the classifiers using different FSS and the number of features yielded statistically same AUC except BN, SVML, SVMS, MLP, NB, and RF. Table 11 presents the comparison of results using *L_T* feature set. SVML, SVMP, SVMR, SVMS, LR, MLP, RF, and CART yielded statistically the same result using all combined 196 features of LIWC and TAALES. Generally, SVM performs well on a large number of features; therefore, in this case, SVM with all four kernels yielded the best results. Table 12 presents the comparison of results using *L_T_D* feature set. J48, JRip, SVML, LMT, LR, MLP, NB, RF, and CART yielded statistically the same result using different FSS.

Tables 13 to 18 present results on AMA dataset with respect to seven feature sets and thirteen classifiers under the 10-FCV framework. Table 13 presents the results of various feature combinations. The LR yielded the best AUC of 88.86% using 200 statistically selected features using GR from D_T feature set and given parameter values. We performed *t*-test among seven FSet over AUC. Some of the t-stat values, which are listed in the last column of Table 13, are less than 2.83. Those t-stat values indicate that the AUC obtained using D, D_L, D_T, L_T, and L_T_D are statistically same. Table 14 presents the comparison of results using D feature set. The LR yielded the best AUC of 85.5% on 60 statistically selected features using GR from D feature set and given parameter values. With the help of *t*-test, we found that BN, SVML, LMT, and RF using GR based 100 features, JRip and LR using GR based 60 features, MLP using TSTAT 80 features, and CART using IG based 100 features yielded statistically the same AUC using tuned parameters. Table 15 presents the comparison of results using D_L feature set. The LR yielded the best AUC of 88.6% on 180 statistically selected features using TSTAT from D_L feature set and given parameter values. The *t*-test indicated that SVML using correlation based 200 features, LMT using IG based 180 features, LR using TSTAT based 180 features, and RF using GR based 200 features yielded statistically the same AUC using tuned parameters. Table 16 presents the comparison of results using D_T feature set. The LR yielded the best AUC of 88.86% on 200 statistically selected features using GR from *D_T* feature set and given parameter values. The *t*-test indicated that BN and LR using GR based 200 features, SVML using TSTAT based 200 features, and LMT using GR based 180

Table 7
Comparison of results for BUR dataset using L_T_D feature set.

FSS	N.F.	Model	Parameter	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F1 (%)
CORR	15	**BN**	Default	75	90.2	82.6	33.9	46.7
IG	300	**J48**	-C 0.2 -M 3	64	99.1	81.5	82.1	71.9
CHI	250	**JRip**	-F 3 -N 1.0 -O 2 -S 1	61	98.6	79.8	74.4	67
IG	250	**SVML**	-C 65,536.0	64	97.9	80.9	66.7	65.3
CHI2	10	**SVMP**	-D 5 -G 2.0 -R 128.0 -C 8.0	60	98.3	79.2	70.6	64.9
CORR	10	**SVMR**	-G 0.25 -C 4096.0	46	99.5	72.7	85.2	59.7
CHI	10	**SVMS**	-G 0.0078125 -R 0.5 -C 32,768.0	48	98.5	73.2	67.6	56.1
TSTAT	300	**LMT**	-I −1 -M 5 -W 0.2	73	98.9	85.9	81.1	76.8
TSTAT	300	**LR**	-R 0.1 -M −1	78	97.9	**88**	71.6	74.6
IG	100	**MLP**	Default	74	96.7	85.3	59.7	66.1
TSTAT	300	**NB**	Default	94	84.3	89.2	28.7	43.9
CHI	10	**RF**	-K 7 -I 110	53	99.1	76.1	80.3	63.9
CORR	20	**CART**	-M 2 -N 5 -C 1.0 -S 1	57	98.5	77.7	71.3	63.3

Table 8
Results of various feature combinations for the FAKE dataset.

FSet	Model	FSS	NF	Model parameters	SPE (%)	SEN (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
D	BN	CORR	50	Default	86.26	86.06	86.16	65.7	74.51	8.2
D_L	LMT	CHI	200	Default	97.87	96.69	**97.28**	93.37	95	**0.2**
D_T	LMT	TSTAT	180	-I –1 -M 5 -W 0.1	97.09	90.82	93.95	90.8	90.81	4.6
L	LMT	CHI	30	-I –1 -M 25 -W 0.15	98.19	96.68	**97.43**	94.42	95.5	–
L_T	LMT	Nil	196	Default	97.87	95.17	**96.52**	93.27	94.2	**0.7**
L_T_D	LMT	CORR	200	Default	98.03	96.18	**97.11**	93.8	95	**0.6**
T	SVML	CORR	100	-C 128.0	97.01	90.58	93.8	90.43	90.5	4.1

Table 9
Comparison of results for FAKE dataset using D_L feature set.

FSS	NF	Model	Parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	10	**BN**	Default	100	86.5	93.2	69.56	82	3.99
CHI	20	**J48**	Default	94.15	97.24	**95.7**	91.36	92.7	**2.02**
CORR	100	**JRip**	Default	95.67	97.17	**96.4**	91.26	93.4	**0.87**
IG	60	**SVML**	-C 8192.0	92.37	97.24	94.8	91.4	91.9	2.92
TSTAT	200	**SVMP**	-D 8 -G 4.88E-4 -R 0.5 -C 128.0	88.83	88.19	88.5	70.4	78.5	9.75
TSTAT	200	**SVMR**	-D 3 -G 0.0039 -C 2.0	85.04	91.5	88.3	75.9	80.2	7.07
CORR	180	**SVMS**	-G 8.63E-5 -R 2.0 -C 3444.31	78.38	93.86	86.1	79.9	79.1	11.2
CHI	200	**LMT**	Default	96.69	97.87	**97.3**	93.37	95	–
TSTAT	40	**LR**	Default	93.64	97.24	**95.4**	91.32	92.5	**1.16**
TSTAT	20	**MLP**	-L 0.02 -M 0.4 -N 1250 -H 20	93.13	96.38	**94.8**	89.3	91.2	**2.63**
IG	10	**NB**	Default	94.15	94.65	94.4	84.47	89	3.02
TSTAT	15	**RF**	-K 7 -I 100	95.94	97.17	**96.6**	91.6	93.7	**0.87**
CORR	160	**CART**	Default	94.66	96.85	**95.8**	90.29	92.4	**1.64**

Table 10
Comparison of results for FAKE dataset using *L* feature set.

FSS	NF	Model	Parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	10	**BN**	Default	89.82	95.28	92.6	85.47	87.6	3.52
CHI	20	**J48**	Default	94.15	97.24	**95.7**	91.36	92.7	**2.51**
CHI	70	**JRip**	Default	95.93	97.17	**96.5**	91.28	93.5	**1.42**
GR	80	**SVML**	-C 512.0	91.87	96.85	94.4	90.33	91.1	4.18
GR	10	**SVMP**	–D 6 -G 2.0 -R 512.0 -C 512.0	96.95	95.67	**96.3**	87.66	92.1	**1.71**
GR	10	**SVMR**	-G 1.0 -C 32,768.0	96.45	96.85	**96.6**	90.77	93.5	**1.21**
GR	10	**SVMS**	-G 0.03125 -R 0.03125 -C 32,768.0	93.13	95.35	**94.24**	86.34	89.6	3
CHI	30	**LMT**	-I –1 -M 25 -W 0.15	96.68	98.19	**97.43**	94.42	95.5	**–**
TSTAT	15	**LR**	-R 1.0E-12	94.66	96.61	**95.6**	89.9	92.2	**2.46**
TSTAT	20	**MLP**	-L 0.02 -M 0.4 -N 1250 -V 20 -H 20	93.13	96.38	94.8	89.27	91.2	3.02
CORR	10	**NB**	Default	93.89	94.8	94.4	84.83	89.1	3.6
TSTAT	15	**RF**	-K 7 -I 100	95.94	97.17	**96.6**	91.56	93.7	**1.16**
CORR	70	**CART**	Default	94.66	96.85	**95.8**	90.29	92.4	**2**

Table 11
Comparison of results for FAKE dataset using *L_T* feature set.

FSS	NF	Model	Parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
Nil	196	**BN**	Default	84.99	87.72	86.35	68.16	75.7	10.6
Nil	196	**J48**	Default	92.37	97.09	94.73	90.75	91.6	14.8
Nil	196	**JRip**	Default	93.64	96.93	95.28	90.42	92	22.7
Nil	196	**SVML**	-C 8.0	92.1	97.48	**94.79**	91.95	92	**0.53**
Nil	196	**SVMP**	-D 8 -G 0.026 -R 0.5 -C 4.0	92.87	97.48	**95.18**	92.08	92.5	**0.15**
Nil	196	**SVMR**	-G 0.0625 -C 32.0	92.36	98.27	**95.31**	94.35	93.3	**–**
Nil	196	**SVMS**	-G 0.0033 -R 0.125 -C 1217.748	90.57	97.56	**94.06**	92.12	91.3	**1.15**
Nil	196	**LMT**	Default	95.17	97.87	96.52	93.27	94.2	13.7
Nil	196	**LR**	-R 1.0	92.61	97.64	**95.12**	92.39	92.5	**0.2**
Nil	196	**MLP**	-L 0.02 -M 0.2 -N 1000 -H 98	91.06	96.77	**93.91**	89.97	90.5	**1.36**
Nil	196	**NB**	Default	90.08	97.01	93.54	90.31	90.2	18.1
Nil	196	**RF**	-K 15 -I 120	88.29	99.53	**93.91**	98.31	93	**1.34**
	196	**CART**	-M 5 -N 5 -C 1.0 -S 1	92.1	97.95	**95.02**	93.41	92.8	**0.35**

features yielded statistically the same AUC using tuned parameters. Table 17 presents the comparison of results using *L_T* feature set. For the L_T feature set, all the classifiers using a different combination of features are statistically significantly same except BN using TSTAT based 60 features, SVMS using correlation based ten features and NB. Table 18 presents the comparison of results using *L_T_*D feature set. The LR yielded the best AUC of 88.6% on 200 statistically selected features using CORR from L_T_D feature set and given parameter values. The *t*-test indicated that JRip using CHI based 200 features, SVMP using CORR based ten features, LMT using TSTAT based 180 features, LR using CORR based 160 features, and RF using GR based 120 features yielded statistically the same AUC using tuned parameters.

Using above-mentioned results, we gained two kinds of insights viz. determining decision rules from classifiers and ascertaining the significance of attributes obtained from an ensemble of three

Table 12
Comparison of results for FAKE dataset using *L_T_D* feature set.

FSS	NF	Model	Parameters	SPE (%)	SEN (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	10	**BN**	Default	86.46	100	93.23	69.56	82	4.35
TSTAT	160	**J48**	Default	97.4	94.15	**95.78**	91.81	93	**2.36**
CORR	120	**JRip**	Default	97.17	95.17	**96.17**	91.22	93.2	**1.46**
CHI	180	**SVML**	-C 8.0	97.8	90.57	**94.18**	92.95	91.7	**2.78**
GR	60	**SVMP**	Default	99.84	26.46	63.15	98.11	41.7	30.1
TSTAT	200	**SVMR**	-G 6.11E-5 -C 8192.0	96.3	90.1	93.2	88.55	89.3	4.31
TSTAT	60	**SVMS**	-G 1.0265E-4 -R 0.03125 -C 13,777.25	97.01	82.69	89.85	89.64	86	7.98
CORR	200	**LMT**	Default	98.03	96.18	**97.11**	93.8	95	**–**
CORR	40	**LR**	Default	98.03	94.66	**96.35**	93.7	94.2	**0.8**
IG	80	**MLP**	-L 0.02 -M 0.6 -N 500 -H 30	97.17	91.84	94.5	91.28	91.6	3.52
CHI	100	**NB**	Default	97.32	92.88	**95.1**	91.48	92.2	**2.38**
CHI	40	**RF**	-K 14 -I 110	98.43	95.43	**96.93**	94.93	95.2	**0.22**
CORR	60	**CART**	Default	96.77	93.89	**95.33**	90	91.9	**2.04**

Table 13
Results of various feature combinations for AMA dataset.

FSet	Model	FSS	NF	Model parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
D	LR	GR	60	-R 1.0E-9	71.3	99.7	**85.5**	97.9	82.5	**1.1**
D_L	LR	TSTAT	180	Default	79.76	97.44	**88.6**	86.15	82.8	**0.19**
D_T	LR	GR	200	-R 1.0E-5	77.97	99.76	**88.86**	98.5	87.04	**–**
L	LR	Nil	93	-R 0.01	62.73	97.4	80.06	83.45	71.22	5.02
L_T	RF	IG	10	-K 7 -I 120	74.76	97.64	**86.2**	86.5	80.2	**1.67**
L_T_D	LR	CORR	160	Default	79.56	97.68	**88.6**	87.25	83.23	**0.18**
T	LMT	Nil	103	Default	47.9	96.12	72.01	71.13	57.25	11.54

Table 14
Comparison of results for AMA dataset using D feature set.

FSS	NF	Model	Parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	100	**BN**	Default	70.7	99.6	**85.2**	97.25	81.9	**0.25**
TSTAT	100	**J48**	-C 0.25 -M 2	55.1	99.4	77.2	94.1	69.5	4.2
GR	60	**JRip**	-F 3 -N 1.0 -O 2 -S 1	65.7	99.9	**82.8**	99.2	79	**1.77**
GR	100	**SVML**	-C 128.0	69.7	99.9	**84.8**	99.2	81.9	**0.39**
CHI	10	**SVMP**	-D 8 -G 4.8828125E-4 -R 32.0 -C 0.03125	53.1	99	76	91.4	67.2	6.24
CORR	10	**SVMR**	-G 0.25 -N 0.5 -C 128.0	51.1	99.1	75.1	92.3	65.8	6.83
CORR	10	**SVMS**	-G 0.125 -R 0.03125 -C 2048.0	52.1	98.8	75.5	89.9	66	6.83
GR	100	**LMT**	-I −1 -M 5 -W 0.1	70.1	99.9	**85**	99.2	82.1	**0.29**
GR	60	**LR**	-R 1.0E-9	71.3	99.7	**85.5**	97.9	82.5	**–**
TSTAT	80	**MLP**	-L 0.18 -M 0.4 -N 500 -V 20 -S 0 -E 20 -H 50	67	97.3	**82.1**	84.1	74.6	**1.61**
TSTAT	100	**NB**	Default	55.3	97.6	76.4	81.9	66	5.86
GR	100	**RF**	-K 7 -I 110	69.9	99.9	**84.9**	99.2	82	**0.35**
IG	100	**CART**	Default	66.3	99.2	**82.7**	94.03	77.8	**1.99**

Table 15
Comparison of results for AMA dataset using D_L feature set.

FSS	NF	Model	Parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	60	**BN**	Default	71.14	99.6	85.4	97.26	82.2	3.18
TSTAT	160	**J48**	Default	70.74	96.56	83.6	80.41	75.3	4.42
GR	40	**JRip**	Default	67.54	99.36	83.5	95.47	79.1	5.04
CORR	200	**SVML**	-C 512.0	75.96	98.6	**87.3**	91.91	83.2	**1.2**
CHI	10	**SVMP**	-D 6 -G 8.0 -R 32.0 -C 32,768.0	36.9	94.76	65.83	59.58	45.6	13.1
GR	10	**SVMR**	-G 4.0 -C 8.0	52.9	99.96	76.4	99.57	69.1	9.37
GR	10	**SVMS**	-G 0.5 -R 0.5 -C 128.0	52.7	99.96	76.33	99.57	68.9	9.33
IG	180	**LMT**	Default	74.75	98.64	**86.7**	91.65	82.3	**1.44**
TSTAT	180	**LR**	Default	79.76	97.44	**88.6**	86.15	82.8	–
TSTAT	120	**MLP**	Default	65.73	95.64	80.7	75.06	70.1	4.77
TSTAT	200	**NB**	Default	60.12	92.11	76.1	60.36	60.2	14.4
GR	200	**RF**	-K 6 -I 90	71.15	99.48	**85.3**	96.59	81.9	**2.38**
GR	120	**CART**	Default	69.34	99.24	84.3	94.79	80.1	8.06

feature sets. To determine the decision rules, we relied on non-black-box approach. Since a black-box approach doesn't yield internal knowledge learned by a classifier. To obtain internal decision rules, we were dependent on decision tree or similar kind of classifiers. As we employed some classifiers like J48, JRip, CART, and LMT, each of them can yield rules for classifying a sample. Out of those four classifiers, we selected a classifier generating informative, useful and the least number of classification rules, so that it will be easier to interpret and use in real life. In addition to the decision rules, we were also interested in a time complexity of the proposed method. Out of various feature sets which yielded statistically same results, we wanted to use the feature set having less number of features. The fewer features will reduce the time complexity of the proposed framework. To determine important fea-

Table 16
Comparison of results for AMA dataset using D_T feature set.

FSS	NF	Model	Parameters	SPE (%)	SEN (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	200	**BN**	Default	99.52	70.94	**85.23**	96.72	81.85	**2.57**
TSTAT	180	**J48**	-C 0.3 -M 3	94.88	64.13	79.51	71.68	67.7	5.49
GR	200	**JRip**	Default	99.88	66.53	83.21	99.1	79.61	3.98
TSTAT	200	**SVML**	-C 2048.0	97.4	73.56	**85.48**	85.35	79.02	**1.95**
CORR	10	**SVMP**	-D 3 -G 3.815E-6 -R 256.0 -C 2.0	77.22	66.75	71.98	36.9	47.53	9.46
GR	10	**SVMR**	-G 4.0 -C 8.0	99.96	52.9	76.43	99.57	69.09	7.74
CORR	10	**SVMS**	-G 0.0093 -R 0.03125 -C 23,170.48	99.32	42.89	71.1	92.72	58.65	12.2
GR	180	**LMT**	-I −1 -M 15 -W 0.0	99.88	71.54	**85.71**	99.18	83.12	**1.89**
GR	200	**LR**	-R 1.0E-5	99.76	77.97	**88.86**	98.5	87.04	**–**
GR	25	**MLP**	Default	99.88	61.92	80.9	99.04	76.2	5.75
CORR	40	**NB**	Default	89.67	67.33	78.5	56.57	61.48	7.2
GR	180	**RF**	-K 6 -I 70	99.88	71.35	**85.61**	99.23	83.01	**2.09**
GR	60	**CART**	-M 2 -N 5 -C 1.0 -S 1	99.88	65.33	82.6	99.12	78.75	3.87

Table 17
Comparison of results for AMA dataset using L_T feature set.

FSS	NF	Model	Parameters	SEN (%)	SPE (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
TSTAT	60	**BN**	Default	78	80.1	79.05	43.99	56.25	4.2
IG	15	**J48**	-C 0.25 -M 3	69.13	96.28	**82.7**	79.18	73.81	2.2
CHI	10	**JRip**	-F 3 -N 5.0 -O 2 -S 1	70.76	95.68	**83.22**	77.18	73.83	1.6
TSTAT	160	**SVML**	-C 512.0	66.73	96.64	**81.68**	80.39	72.93	2.7
TSTAT	50	**SVMP**	-D 4 -G 0.42 -R 1.41 -C 13.45	68.74	95.08	**81.91**	73.84	71.2	2.5
CORR	80	**SVMR**	-G 0.0625 -C 32,768.0	69.75	96.32	**83.03**	79.31	74.22	1.9
CORR	10	**SVMS**	-G 0.02 -R 0.03125 -C 46,340.95	28.88	98.4	63.64	77.79	42.12	13
CHI	10	**LMT**	-I −1 -M 25 -W 0.15	66.97	97.36	**82.16**	83.64	74.38	2.2
TSTAT	180	**LR**	-R 0.1	68.55	96.76	**82.65**	80.95	74.24	2
CHI	15	**MLP**	-L 0.02 -M 0.2 -N 1250 -V 20 -S 0 -E 20 -H 10	61.66	97.52	**79.59**	83.42	70.91	**2**
Nil	196	**NB**	Default	60.32	90.35	75.33	55.54	57.83	8.13
IG	10	**RF**	-K 7 -I 120	74.76	97.64	**86.2**	86.5	80.2	**–**
IG	15	**CART**	Default	70.74	96.16	**83.45**	78.62	74.47	**1.96**

Table 18
Comparison of results for AMA dataset using L_T_D feature set.

FSS	NF	Model	Parameters	SPE (%)	SEN (%)	AUC (%)	PREC (%)	F_1 (%)	t-stat
GR	60	**BN**	Default	99.6	71.14	85.4	97.26	82.17	3.49
TSTAT	180	**J48**	Default	95.76	72.75	84.3	77.4	75	3.85
CHI	200	**JRip**	Default	95.52	76.95	**86.2**	77.42	77.18	**1.99**
CORR	200	**SVML**	-C 32,768.0	96.8	77.16	**87**	83.21	80.07	**1.35**
CORR	10	**SVMP**	-D 9 -G 9.0 -R 2.0 -C 1722.156	97.08	78.17	**87.6**	84.56	81.24	**0.88**
GR	10	**SVMR**	-G 4.0 -C 8.0	99.96	52.9	76.4	99.57	69.09	9.87
CORR	10	**SVMS**	-G 0.022 -R 0.03125 -C 46,340.95	98.4	28.88	63.6	77.79	42.12	16.8
TSTAT	180	**LMT**	-I 8 -M 10 -W 0.2	97.28	77.54	**87.4**	85.33	81.25	**1.02**
CORR	160	**LR**	Default	97.68	79.56	**88.6**	87.25	83.23	**–**
CORR	10	**MLP**	Default	97.16	69.54	83.4	83.01	75.68	3.03
IG	200	**NB**	Default	91.51	61.92	76.7	59.31	60.59	10.7
GR	120	**RF**	-K 6 -I 80	99.44	71.35	**85.4**	96.34	81.98	**2.45**
GR	100	**CART**	Default	99.24	69.34	84.3	94.79	80.09	4.27

tures from an ensemble of three feature sets, we relied on odds ratio yielded by LR. The odds ratios indicated the significance of features out of different feature set.

To differentiate a feature of LIWC or TAALES from a term of DTM, we added a suffix as 'L' and 'T' for each feature of LIWC and TAALES respectively. For BUR dataset, LR yielded 6.26% more AUC than LMT and LMT outperformed the rest of classifiers. Moreover, one of the interesting parts of LMT is that it drops irrelevant features for the classification task automatically. Therefore, we chose LMT to report the decision rules. LMT generated only one classification rule using IG based 250 features of D_L feature set, which is presented in Fig. 2. Since LMT yielded only one classification rule, so it had only one node i.e. root node. The rule is the multivariate linear regression equation fitted on the root node. In the given rule, the intercept, features, and coefficients are depicted. All features and coefficients represent variables and parameter values of the linear regression respectively. The presented rule for a given class has positive and negative coefficients. The positive and negative coefficients represent the positive and negative relationship with the corresponding class. Furthermore, the higher the coefficient value more is its impact on obtaining the probability for the respective class. In Fig. 2, 'Class 1' represents the positive class. Out of 250 features, LMT is dependent only on 93 features, so the rest of features are not contributing to classification. According to the positive class rule and the second best result, only a few features of LIWC have been found important for satire detection namely *ExclamL, swearL, iL, QuoteL, QMarkL, informalL, youL, certainL, weL, focusfutureL*, and *theyL*. If we exclude the features related to punctuation marks, the rest of features can be considered under generic features for satire detection. For FAKE dataset, *D_L, L, L_T,* and *L_T_D* yielded statistically the same results. CHI-based 30 features obtained from L feature set were processed by LMT to select 19 relevant features for fitting the model. LMT yielded only one rule for classification, which is a special case of LMT having a single node. The decision rule is presented in Fig. 3. The rule indicates that the distribution of some features like

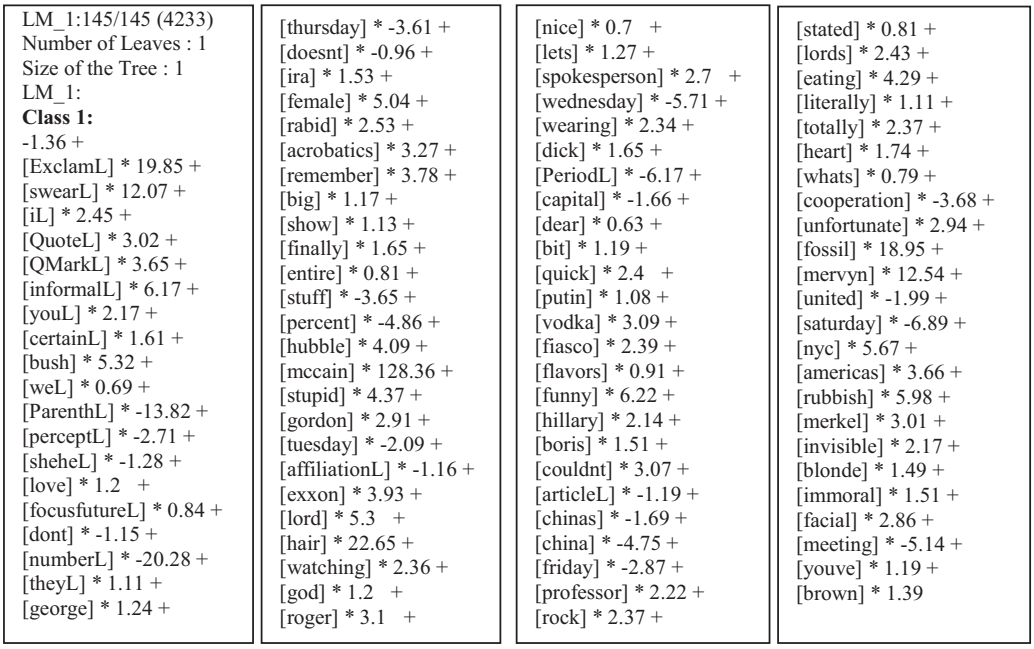

Fig. 2. Decision rules for satire detection on BUR dataset using DL, IG, 250 and LMT.

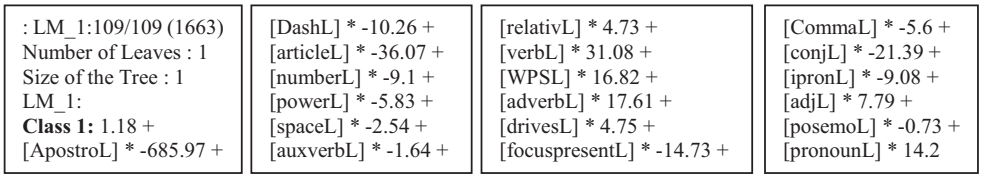

Fig. 3. Decision rules for satire detection on the FAKE dataset using CHI, 30, L, and LMT.

ApostroL, DashL, articleL, numberL, powerL, spaceL, auxverbL, focuspresentL, CommaL, conjL, ipronL, and *posemoL,* are positively associated with the negative class i.e. the serious news. On the other hand, the distribution of some features like *relativL, verbL, WPSL, adverbL, drivesL, adjL, and pronounL,* are positively associated with the positive class i.e. the satiric news. For AMA dataset, five feature sets produced statistically the same result. Using GR based 120 features of D_L, CART yielded one of the informative and easily interpretable and presentable rules as presented in Fig. 4. According to this rule, we found that only two LIWC features are important for irony detection i.e. *sexualL* and *deathL.*

In addition to decision rules, we highlighted the significance of features obtained from the ensemble of three feature sets for each dataset. Regarding this, we considered the output of LR on *L_T_D* feature set to explain the significance of different features. In linear regression model, if a feature is useful in predicting the output variable, mean-square regression (MSR) should be large, thereby making the statistic $F=$ (MSR/Mean Squared Error) also large. We determine the significance of a predictor variable in LR in a different manner. The log-odds ratio reported by LR for each dataset is quite important to understand the relationship between a predictor and the class variable. By exponentiating log-odds ratio, we obtain odds ratio. A one-unit change in a predictor variable leads to a multiplicative change in the odds of the odds ratio. In other words, the odds ratio gives the multiplicative increase in odds by a unit of change in the predictor variable. The negative log-odds ratio will yield a negative relationship with the probability of class variable of interest and the independent variable; positive log-odds ratio will indicate a positive relationship. Regarding odds ratio, the relationship of a predictor with class variable depends on the value of odds ratio. The odds ratio less than 1 indicates negative relationship, more than 1 indicates positive relationship, and 1 indicates no relationship. Furthermore, the higher the odds ratio more is its impact on the class variable, hence can be considered as a better predictor.

As regards BUR dataset, out of the TSTAT based features of *L_T_D*, only those having odds ratio more than 1, numbering 107, are presented in Table 19. These 107 features are positively related with satiric news, thereby implying that they can play a major role in satiric news detection. We found out that 20 LIWC features had a positive relationship with *satiric* news. Out of them, *swearL, posemoL, bodyL, negemoL, seeL, sexualL, focusfutureL, rewardL, focuspresentL, maleL, certainL,* and *leisureL* are psychological features of satiric content. Further, 36 features obtained from TAALES have a positive relationship with *satiric* news. Out of thirty-six TAALES features, *Brysbaert.Concreteness.Unigram.CWT, MRC.Familiarity.CWT, MRC.Concreteness.CWT, MRC.Meaningfulness.AWT, MRC.Concreteness.FWT,* and *MRC.Meaningfulness.FWT* are psycholinguistic features. TAALES support four kinds of psycholinguistic features viz. *Familiarity, Imageability, Concreteness, and Meaningfulness*. Out of these four, *imageability* was missing in the given corpus. The rest of features are *word frequency indices* with respect to All Words (AW), Content Words (CW), and Function Words (FW). Similarly, there are 51 DTM features, whose odds ratio were more than 1 and had a positive relationship with *satiric* news. For FAKE dataset, out of the correlation based 16 features of *L_T_D,* only those having odds ratio more than 1, are presented in Table 20. There are 14 TAALES features, which are negatively associated with serious news. Out of those 14 features, *MRC.Imageability.FWT, MRC.Meaningfulness.FWT, and MRC.Meaningfulness.AWT* are three

```
banana < 0.0416665
|  milk < 0.0333335
|  |  wolf < 0.0555555
|  |  |  bic < 0.125
|  |  |  |  uranium < 0.25
|  |  |  |  |  women < 0.0833335
|  |  |  |  |  |  sexualL < 0.124389
|  |  |  |  |  |  |  wolves < 0.0454545
|  |  |  |  |  |  |  |  head < 0.1666665
|  |  |  |  |  |  |  |  |  bananas < 0.0625
|  |  |  |  |  |  |  |  |  |  woman < 0.25
|  |  |  |  |  |  |  |  |  |  |  powers < 0.25
|  |  |  |  |  |  |  |  |  |  |  |  deathL < 0.1986755
|  |  |  |  |  |  |  |  |  |  |  |  |  loss < 0.25
|  |  |  |  |  |  |  |  |  |  |  |  |  |  tuscan < 0.05
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  ladies < 0.1666665
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  skin < 0.1666665
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  walked < 0.5
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  damn < 0.5
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  stomach < 0.5
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  complete < 0.25
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  gender < 0.125
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sound < 0.25
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  feminine < 0.25
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  cream < 0.5
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  town < 0.5
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  neighbors < 0.25: neg(2482.0/128.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  neighbors >= 0.25: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  town >= 0.5: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  cream >= 0.5: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  feminine >= 0.25: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  sound >= 0.25: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  gender >= 0.125: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  complete >= 0.25: pos(2.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  stomach >= 0.5: pos(3.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  damn >= 0.5: pos(3.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  walked >= 0.5: pos(3.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  skin >= 0.1666665: pos(3.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  ladies >= 0.1666665: pos(3.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  tuscan >= 0.05: pos(3.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  loss >= 0.25: pos(4.0/0.0)
|  |  |  |  |  |  |  |  |  |  |  |  deathL >= 0.1986755: pos(8.0/6.0)
|  |  |  |  |  |  |  |  |  |  |  powers >= 0.25: pos(5.0/0.0)
|  |  |  |  |  |  |  |  |  |  woman >= 0.25: pos(6.0/1.0)
|  |  |  |  |  |  |  |  |  bananas >= 0.0625: pos(7.0/0.0)
|  |  |  |  |  |  |  |  head >= 0.1666665: pos(9.0/0.0)
|  |  |  |  |  |  |  wolves >= 0.0454545: pos(10.0/0.0)
|  |  |  |  |  |  sexualL >= 0.124389: pos(20.0/7.0)
|  |  |  |  |  women >= 0.0833335: pos(16.0/0.0)
|  |  |  |  uranium >= 0.25: pos(19.0/0.0)
|  |  |  bic >= 0.125: pos(32.0/1.0)
|  |  wolf >= 0.0555555: pos(33.0/0.0)
|  milk >= 0.0333335: pos(83.0/1.0)
banana >= 0.0416665: pos(87.0/0.0)
Number of Leaf Nodes: 28
Size of the Tree: 55
```

Fig. 4. Classification rules for irony detection on AMA dataset using GR, 120, DL, and CART.

psycholinguistic features, which helped to discriminate non-satiric and satiric news in the Indian context. There are two LIWC features *AnalyticL and SixltrL*, which are positively related to satiric news. For AMA dataset, out of the correlation based features of L_T_D, only those having odds ratio more than 1, numbering 94, are presented in Table 21. There are 44 features obtained from LIWC, which are positively related to satiric reviews. Out of 44 features, *relativL, bodyL, femaleL, hearL, maleL, sexualL, ingestL, insightL, homeL, affiliationL, tentatL, healthL, seeL, socialL, netspeakL, negemoL, focuspastL, differL, deathL, causeL, riskL, anxL, angerL, swearL, informalL, achieveL, leisureL, certainL, and interrogL* are psychological features. There are 18 features obtained by TAALES, which have positive relationships with *ironic* reviews. Out of 18 features, *Brysbaert.Concreteness.Unigram.FWT, MRC.Imageability.FWT, MRC.Concreteness.FWT, and Brysbaert.Concreteness.bigrams.- AWT* are psycholinguistic features. There are 32 features obtained from DTM, which are positively related to ironic reviews.

6. Human evaluation

We performed the human evaluation in order to realize the difficulty involved in automatic detection of satire and irony found in news and customer reviews. We considered around 25% samples of FAKE and AMA datasets using stratified random sampling

Table 19
The odds ratio with respect to positive class for BUR dataset.

Feature name	O.R.[a]	Feature name	O.R.	Feature name	O.R.
BNC.Written.Range.AWT	Infinity	BNC.Spoken.Range.AWT	5.13×10^{24}	rightwing	2,118,122.9
DicL	Infinity	ceasefire	6.13×10^{23}	bilateral	820,844.15
TL.Freq.AW.LogT	3.94×10^{205}	developing	1.67×10^{22}	conjL	382,201.43
SUBTLEXus.Freq.CW.LogT	2.14×10^{199}	Kuperman.AoA.FWT	1.38×10^{22}	talks	222,027.22
pronounL	1.32×10^{176}	south	1.43×10^{21}	official	75,870.28
Kuperman.AoA.CWT	1.16×10^{171}	development	3.62×10^{20}	arrive	8225.33
SUBTLEXus.Range.AW.LogT	1.14×10^{170}	TL.Freq.CWT	2.98×10^{20}	focuspresentL	5704.19
Brysbaert.Concreteness.Unigram.CWT	3.64×10^{137}	peacekeeping	2.44×10^{18}	staged	2247.16
MRC.Familiarity.CWT	1.38×10^{131}	gains	3.80×10^{17}	turnover	1606.07
KF.Freq.FW.LogT	2.61×10^{106}	hong	2.46×10^{16}	sundays	989.36
auxverbL	4.58×10^{104}	infrastructure	1.58×10^{16}	maleL	907.2
MRC.Concreteness.CWT	3.13×10^{97}	hosni	2.57×10^{15}	weL	789.97
swearL	4.14×10^{93}	total	4.54×10^{13}	pakistani	780.4
SUBTLEXus.Range.CWT	4.42×10^{82}	bosnian	2.86×10^{13}	government	360.5
BNC.Spoken.Range.CWT	2.50×10^{81}	african	8.08×10^{12}	province	248.29
KF.Freq.AWT	6.75×10^{67}	sector	6.45×10^{11}	rebels	234.41
BNC.Spoken.Freq.FWT	3.55×10^{67}	MRC.Meaningfulness.FWT	2.60×10^{11}	pakistan	206.24
KF.Freq.FWT	4.92×10^{66}	exchange	2.58×10^{11}	implementation	205.54
KF.Nsamp.CWT	7.07×10^{60}	regional	1.25×10^{11}	twoday	139.65
MRC.Meaningfulness.AWT	3.42×10^{60}	focusfutureL	9.36×10^{10}	certainL	107.7
KF.Ncats.AWT	2.94×10^{57}	argentina	2.22×10^{10}	secure	99.46
index	1.05×10^{57}	clashes	2,362,620,458	tension	93.3
SUBTLEXus.Freq.AWT	3.26×10^{52}	BNC.Spoken.Trigram.ProportionT	2,000,178,710	ties	78.84
palestinian	9.61×10^{46}	rivals	1,230,610,137	leisureL	60.12
BNC.Written.Bigram.Freq.NormedT	1.75×10^{44}	sexualL	1,046,328,385	BNC.Spoken.Freq.CW.LogT	58.03
adverbL	4.00×10^{42}	tuesday	718,834,036	combat	45.6
Brown.Freq.AW.LogT	8.89×10^{35}	yuan	507,639,198	negateL	24.6
MRC.Concreteness.FWT	2.23×10^{34}	million	153,183,328	Nationalist	19.77
SUBTLEXus.Range.FWT	1.36×10^{34}	Brown.Freq.FWT	152,164,197	Lebanon	13.32
posemoL	2.10×10^{33}	yen	43,923,175.8	Meeting	11.11
QuoteL	1.11×10^{30}	AWL.Sublist.1.NormedT	22,179,137.6	China	7.15
bodyL	5.42×10^{29}	AWL.Sublist.6.NormedT	13,559,159.9	AWL.Sublist.4.NormedT	2.38
investment	8.54×10^{27}	fidel	12,754,192.4	Framework	1.76
BNC.Written.Freq.FW.LogT	8.15×10^{25}	unchanged	9,271,495.79	BNC.Spoken.Trigram.Normed.LogT	1.6
negemoL	4.85×10^{25}	rewardL	6,268,642.3	BNC.Written.Freq.CWT	1.37
seeL	8.60×10^{24}	AWL.Sublist.2.NormedT	2,147,869.47		

a. O.R. stands for Odds Ratio

Table 20
The odds ratio with respect to positive class for FAKE dataset.

Feature name	O.R.	Feature name	O.R.
KF.Freq.FW.LogT	9.47×10^{61}	BNC.Written.Freq.AWT	295,892,761
SUBTLEXus.Freq.FWT	4.71×10^{41}	BNC.Spoken.Bigram.ProportionT	22,759,780.2
BNC.Spoken.Freq.FW.LogT	4.02×10^{40}	BNC.Spoken.Bigram.NormedT	3,226,751.54
SixltrL	4.81×10^{32}	MRC.Imageability.FWT	1,983,788.18
TL.Freq.FWT	9.16×10^{29}	MRC.Meaningfulness.FWT	33,270.266
KF.Freq.AWT	1.25×10^{28}	KF.Nsamp.FWT	5.2373
BNC.Spoken.Freq.AWT	3.42×10^{12}	MRC.Meaningfulness.AWT	3.513
AnalyticL	1.21×10^{11}	BNC.Written.Bigram.Freq.Normed.LogT	2.5038

for human evaluation. Accordingly, we selected 100 positive and 323 negative samples from FAKE dataset. Further, 125 positive and 625 negative samples were selected from AMA dataset. For both datasets, the ratio of positive and negative samples was maintained as in original datasets. We couldn't employ annotators on whole datasets because we conjecture that annotators may not perform well on repetitive kind of difficult tasks for a long time. According to our experience, annotators admitted that they found it quite difficult to figure out the clues of irony and satire. In order to overcome this difficulty, we employed six annotators, divided into two groups where one group worked on one dataset, while the other was assigned the second dataset. In addition to that, we asked all annotators not to give more than two hours a day for this task in order to not make it look like a tedious job for them. For annotation purpose, we shuffled positive and negative samples together for both datasets. We asked them to reply two questions related to each news/review: (i) Is the given news/review is ironic? (ii) If yes, then indicate the clue of the irony. The inter-annotator agreement among three annotators is reported using Fleiss' Kappa coefficients (κ) [65]. The κ values for irony detection in FAKE and AMA dataset were 0.685 and 0.362 respectively. Accordingly, the agreement level among all annotators for FAKE and AMA dataset were *substantial* and *fair* respectively [66]. The selected news and reviews samples were classified with the help of annotators as well as our proposed approach. To classify a sample using 3 annotators, we employed simple majority voting. We partitioned FAKE and AMA dataset under 10 folds to report the results. We considered the same set of partitions of samples for all classifiers as well as human annotators. We calculated an average of 10 folds of results yielded by human annotators and classifiers. The classification results under 10-FCV are presented in Tables 22 and 23 for FAKE and AMA dataset respectively. For both datasets, we performed *t*-test between human annotators versus different feature set combinations at 1% level of significance and 18 degrees of freedom. Table 22 indicates that all feature sets yielded statistically significantly the same result that compared

Table 21
The odds ratio with respect to positive class for AMA dataset.

Feature name	O.R.	Feature name	O.R.	Feature name	O.R.
Banana	7.40×10^{205}	Shirt	3555.5747	wearing	545.3326
Hutzler	2.17×10^{180}	Back	3451.2718	swearL	484.0206
Bananas	8.95×10^{152}	affiliationL	2.18×10^{11}	woman	357.0135
Milk	9.00×10^{141}	SUBTLEXus.Range.FW.LogT	1.43×10^{11}	man	293.8963
iL	5.16×10^{100}	Bic	1.11×10^{11}	friend	208.1255
Tuscan	1.18×10^{92}	Ore	4.65×10^{10}	star	176.7333
Women	6.32×10^{91}	tentatL	4.47×10^{10}	MRC.Imageability.FWT	162.8794
Wolf	9.65×10^{64}	healthL	3.44×10^{10}	informalL	161.6808
relativL	5.46×10^{58}	seeL	3.25×10^{10}	achieveL	115.6409
prepL	3.66×10^{41}	socialL	6,567,115,349	wife	80.0698
Uranium	2.59×10^{30}	netspeakL	481,783,158	eyes	76.7794
bodyL	2.06×10^{27}	Moon	123,125,470	adverbL	33.2743
WPSL	1.77×10^{27}	BNC.Written.Freq.FW.LogT	24,284,389.4	life	23.5931
pronounL	8.69×10^{26}	negemoL	22,091,932.9	make	10.9774
Head	2.58×10^{26}	focuspastL	10,980,124	TL.Freq.CWT	9.0396
femaleL	1.14×10^{26}	BNC.Spoken.Range.FWT	3,546,075	hours	6.4456
Complete	5.40×10^{22}	differL	2,735,986.5	BNC.Written.Bigram.Freq.Normed.LogT	5.3942
hearL	2.36×10^{21}	deathL	1,054,422	leisureL	3.6641
auxverbL	8.59×10^{19}	World	331,157.9	articleL	3.4745
maleL	1.88×10^{19}	SUBTLEXus.Freq.FW.LogT	37,790.7	certainL	3.3519
weL	1.20×10^{18}	KF.Freq.FW.LogT	33,099.98	BNC.Spoken.Trigram.Normed.LogT	2.9958
AllPuncL	1.33×10^{17}	causeL	31,190.21	home	2.803
CloutL	3.38×10^{16}	negateL	2799.3	MRC.Concreteness.FWT	2.5509
sexualL	8.05×10^{13}	compareL	1923.1	days	2.043
ToneL	4.64×10^{13}	anxL	1787.15	left	1.7977
ingestL	4.29×10^{13}	night	1392.0369	Brysbaert.Concreteness.bigrams.AWT	1.1667
youL	2.01×10^{13}	QMarkL	1169.28	interrogL	1.0774
Gallon	3.92×10^{12}	face	900.99	BNC.Written.Trigram.Freq.Normed.LogT	1.067
insightL	8.29×10^{11}	TL.Freq.AWT	776	SUBTLEXus.Freq.FWT	1.0343
homeL	6.65×10^{11}	Brown.Freq.FW.LogT	752	BNC.Spoken.Trigram.NormedT	1.0327
SUBTLEXus.Freq.AWT	23,144.07	Brysbaert.Concreteness.Unigram.FWT	627.27		
riskL	8518.2	angerL	595.87		

Table 22
The classification results for FAKE dataset under 10-FCV.

Fset	Model	FSS	NF	Model parameters	SEN	SPE	AUC	PRE	F1	t-stat (AUC)
D	LR	GR	250	-R 0.1 -M 15	96	98.15	97.07	94.55	94.98	**2.04**
D_L	LR	GR	160	-R 0.01 -M 20	97	98.17	97.59	94.85	95.65	**2.21**
D_T	LR	GR	250	-R 0.1 -M 15	96	98.15	97.07	94.55	94.98	**2.04**
L	LMT	IG	60	-I −1 -M 5 -W 0.15	96	97.24	96.62	92.07	93.69	**1.54**
L_T	LMT	CHI	140	-I 7 -M 10 -W 0.15	98	97.53	97.77	93.03	95.23	**1.74**
L_T_D	LR	GR	160	-R 0.01 -M 20	98	98.48	98.24	95.61	96.61	**2.39**
T	BN	CORR	10	Default	38	81.73	59.87	39.18	38.58	7.29
–	**Human**	–	–	–	**80**	98.57	**89.29**	98.57	86.04	–

Table 23
The classification results for AMA dataset under 10-FCV.

Fset	Model	FSS	NF	Model parameters	SEN	SPE	AUC	PRE	F1	t-stat (AUC)
D	SVMR	GR	250	-G 8.0 -C 2.0	88.78	100	94.39	100	93.85	3.7
D_L	SVMR	GR	250	-G 2.0 -C 512.0	93.53	99.84	96.69	99.23	96.03	4.5
D_T	SVMR	GR	300	-G 8.0 -C 2.0	91.22	99.84	95.53	99.17	94.81	4.13
L	SVML	GR	30	-G 0.0 -C 65,536.0	85.58	98.4	**91.99**	92.71	88.16	**2.52**
L_T	J48	CORR	40	-C 0.25 -M 2	84.8	99.52	**92.16**	97.25	90.6	**1.83**
L_T_D	SVMR	GR	300	-G 8.0 -C 128.0	95.13	99.68	97.41	98.46	96.58	5.13
T	SVMP	TSTAT	103	-D 5 -G 4.0 -R 2.4E-4 -C 11.3	55.9	87.05	71.47	45.99	49.93	4.49
	Human	–	–	–	**73.85**	95.52	**84.68**	76.59	74.71	–

to human evaluators, whereas human annotators outperformed T. Furthermore, Table 23 helps us infer that human annotators are statistically significantly same as *L* and *L_T. D, D_L, D_T*, and *L_T_D* statistically significantly outperformed human evaluators. Here, we presented some of the reviews, which were not correctly classified by annotators:

‘‘The shirt itself is amazing, no need to review the awesome happening on the front. The shirt is probably the worst cotton made and a large is for someone around 250 pounds. Get a size down when ordering. Doesn’t shrink either after washing. Or feel any better on your body.’’ - Mountain Three Short Sleeve T-Shirt

‘‘I have been using the banana slicer for nearly two years now. I have fond memories of Model 1 A, an ingenious, if clunky (well, more "cardboardy", really) version of these new slicked up models. I have to say, I love the new upgrades on this model over the 570, or even the 571A. Really streamlined the curves and sharpened the edges. The

engineering is very precise, down to the smallest detail of removing the bananger (black booger on the end of the banana) without much loss of banana. They have taken care of most of the safety issues as well, but one must always be reminded "throw the peel in the trash bin BEFORE you begin to slice". I can't wait to see tomorrow's model, which is purported to have the jazzed up name of "572Z", and may slice longitudinally for ice cream enthusiasts!'' - Hutzler 571 Banana Slicer

''I have determined that my children are weak. They need Tuscan milk. Aged Tuscan milk that spends 3--5 days in transport. If you don't have to chew it, it's milk for weaklings.'' - Tuscan Dairy Whole Vitamin Gallon

''Only comes in medium. I always buy my women products in super, super plus, or regular...I gave the pack to my husband to return, since he drives, makes all purchases, and I cannot speak to anyone other than him.'' - BIC Cristal For Her Ball Pen

''I always wondered where Saddam Hussein got his weapons of mass destruction...''- Images SI Inc Uranium Ore

In brief, some of the key observations of the study can be listed as follows:

- The detection of satire in the news is easier than that of irony in customer reviews for humans as well as our proposed approach.
- Our proposed approach yielded the same performance as that of human annotators for satire detection, but better performance for irony detection.
- We can attribute the worse performance of human evaluators compared to our proposed approach on AMA dataset to the lack of knowledge of context and domain. They failed to predict the category of samples because we shuffled ironic and non-ironic customer reviews from ten products together. Therefore, human annotators didn't know which product a customer review was written on. In some cases, ironic utterance depends upon the context and domain, which can be observed in above-mentioned incorrectly classified ironic reviews on "BIC Cristal For Her Ball Pen" and "Images SI Inc Uranium Ore."
- For news corpora like BUR and FAKE, some psycholinguistic features obtained using TAALES are among the best features according to the odds-ratio values. This indicated that a single unit change in these features would affect the decision of satire severely.
- For AMA dataset, some LIWC features are having more odds-ratios than TAALES features, which means that the features obtained using LIWC are better discriminators compared to TAALES features for ironic customer reviews detection.
- For all three datasets, we employed a host of classifiers on 500 or fewer features ensembled from *L*, *T* and *D*. For BUR dataset, 500 features yielded the best performance. Even though 500 features make the input dataset a high dimensional dataset, the classifiers yielded good performances. On the basis of the experiments carried out, we inferred three reasons, which avoided the curse of dimensionality. First, we considered the ridge regularization in LR, which is a good solution to avoid the curse of dimensionality. Table 5 indicates that when we had increased the value of Ridge from 10^{-8} to 10^{-1} for regularization, the performance of LR improved a lot. Increasing in the value of ridge saves the model from overfitting and reduces the variance [44]. Second, introducing more number of relevant features to the dataset helps the model overcome bias [67]. Here, ensembling of features from *L*, *T* and *D* helped us obtaining more number of relevant features. We observed from experiments that having more number of features were not creating any issue instead it was increasing the performance of the classifier. Third, SVM performs quite well on a high dimensional dataset.
- In terms of annotation, annotators reported that indicating the clue of irony or satire is a quite complex task.

7. Conclusion and future directions

In this study, we employed unigrams, semantic, psycholinguistic, and statistical features of the text, to detect satiric and ironic content in news and reviews respectively. Even though standalone LIWC doesn't consider irony, sarcasm, and context [37], our proposed methodology shows that LIWC is proved to be useful for satire and irony detection, if it is used along with other word frequency indices and psycholinguistic features obtained using TAALES, unigrams features of DTM, and statistical feature subset selection method. We ensembled three feature sets generated using DTM, LIWC, and TAALES to determine satire and irony in news and reviews respectively. For the satiric news of BUR dataset, logistic regression yielded the best AUC of 92.07% using t-statistic based 500 ensembled features of DTM, TAALES and LIWC. For the satiric news of FAKE dataset, the LMT yielded the best AUC of 97.43% using χ^2-statistics based 30 features of LIWC. For ironic reviews of AMA dataset, the logistic regression yielded the best AUC of 88.86% using gain-ratio based 200 ensembled features of DTM and TAALES. Some of the common psychological features of LIWC found among BUR and AMA datasets are *bodyL, negemoL, seeL, sexualL, swearL, maleL, certainL,* and *leisureL*. In terms of linguistic features, BUR and AMA share common function words like *pronounL, auxverbL, advL,* and *weL*. By these observations, we can conclude that satire and irony share some common psycholinguistic characteristics obtained using LIWC. On the other hand, TAALES also yielded some psycholinguistic features of the text like *concreteness, meaningfulness, imageability*, and *familiarity*. With respect to BUR and FAKE datasets, they share common psycholinguistics characteristics like *meaningfulness*. Similarly, BUR and AMA have common psycholinguistics characteristics like *concreteness*. Further, the *imageability* is found in satire as well irony as we observed while experimenting with FAKE and AMA datasets. Hence, these experiments strengthen a belief that satire is a sub-type of irony and share some psycholinguistic characteristics. The *t*-test helped us draw some conclusions with respect to feature set and classification techniques. In terms of three feature sets, the output of LIWC turned out to be the most important feature set. Among ten classifiers, we found that logistic regression and random forest to be consistently equivalent by yielding the best classification results for any ensembled feature set of FAKE and AMA datasets.

Although, we proposed a generic approach to detect ironic and satiric content available in reviews and news respectively, there is still a lot of scope for improvement. We considered stylometrics, semantic metrics, and psychometrics of the corpus to determine the satiric and ironic content. In order to improve upon the obtained results, we can further include metrics related to syntactic and pragmatic information as part of our future work. The syntactic information metrics include the distribution of POS including their order of appearances, the distribution of particular phrases, etc. The pragmatic information metrics involves the distribution of intents related words, type of intentions, etc. The intent related

words indicate the intention behind used words in the text. The conceptual level features can also improve upon the obtained results. The conceptual level feature refers to a couple of words, which often occur together like “what you”, “you VRB”, “I VRB”, “as JJ as”, etc. Some examples of “as JJ as” are “as smart as”, “as intelligent as”, “as sharp as”, etc. Determination of target of satire in a news corpus is a quite difficult task for human beings, which can be automated in future. Finally, determination of the degree of satire will be another interesting task related to irony handling that needs to be tackled in future work. Further, Kernel Binary Quantile Regression (KBQR), Differential Evolution trained KBQR (DEKBQR) [68] for classification and Elitist Quantum-Inspired Differential Evolution (QDE) [69] wrapper for feature selection can be attempted in future for obtaining better results.

Acknowledgments

We would like to thank six annotators namely B. Shiva Krishna (B. Tech. CSE), Kumar Mani (MBA), Yogesh Khandelwal (M. Tech. CSE), Lamminthang Singsit (M. Tech. CSE), Gutha Jaya Krishna (M. Tech. AI), and Mulagala Sandhya (M.Tech. CSE) for spending their quality time for annotation in a very short notice.

References

[1] R. Gibbs, H. Colston, Irony in language and thought: A cognitive science reader, 2007.
[2] R. Kreuz, R. Roberts, On satire and parody: The importance of being ironic, Metaphor Symbol (1993).
[3] A. Reyes, P. Rosso, On the difficulty of automatically detecting irony: beyond a simple case of negation, Knowl. Inf. Syst. 40 (2014) 595–614.
[4] K.E. Beckson, A. Ganz, Literary Terms: A Dictionary, Macmillan, 1989.
[5] A. Preminger, F. Warnke, O. Hardison Jr, Princeton encyclopedia of poetry and poetics, 2015.
[6] M. Walker, J. Tree, P. Anand, R. Abbott, J. King, A corpus for research on deliberation and debate, LREC (2012).
[7] S. Skalicky, S. Crossley, A statistical analysis of satirical Amazon. com product reviews, Eur. J. Humour Res. 2 (2015) 66–85.
[8] C. Burfoot, T. Baldwin, Automatic satire detection: Are you having a laugh? in: Proceedings of the ACL-IJCNLP 2009 Conference, 2009, pp. 161–165.
[9] G. Hirsch, S. Blum-Kulka, Identifying irony in news interviews, J. Pragmatics 70 (2014) 31–51.
[10] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, Sarcasm as contrast between a positive sentiment and negative situation, EMNLP (2013) 704–714.
[11] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowl. Based Syst. 89 (2015) 14–46.
[12] Y. Hao, T. Veale, An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes, Minds Mach. 20 (2010) 635–650.
[13] K. Ravi, V. Siddeshwar, V. Ravi, L. Mohan, Sentiment analysis applied to educational sector, in: IEEE Intl. Conf. on Computational Intelligence and Computing Research, IEEE, Madurai, 2015, pp. 117–122.
[14] R. Ghosh, K. Ravi, V. Ravi, A novel deep learning architecture for sentiment classification, in: 3rd IEEE International Conference on Recent Advances in Information Technology, IEEE, ISM, Dhanbad, 2016, pp. 511–516.
[15] K. Ravi, V. Ravi, C. Gautam, Online and semi-online sentiment classification, in: IEEE Intl. Conf. on Computing, Communication & Automation, IEEE, New Delhi, 2015, pp. 925–930.
[16] K. Ravi, V. Ravi, Sentiment classification of Hinglish text, in: 3rd IEEE International Conference on Recent Advances in Information Technology (RAIT2016), ISM, Dhanbad, 2016, pp. 641–645.
[17] A. Reyes, P. Rosso, Making objective decisions from subjective data: detecting irony in customer reviews, Decis. Support Syst. 53 (2012) 754–760.
[18] R. Gibbs, Irony in talk among friends, Metaphor Symbol 15 (2000) 5–27.
[19] P. Grice, Logic and conversation, in: P. Cole, J. Morgan (Eds.), Syntax and Semantics, Academic Press, New York, 1975, pp. 41–58, doi:10.1111/j.1365-2664.2006.01229.x.
[20] D. Wilson, The pragmatics of verbal irony: echo or pretence? Lingua 116 (2006) 1722–1743.
[21] R. Giora, Literal vs. figurative language: Different or equal? Journal of Pragmatics 34 (2002) 487–506.
[22] S. Kumon-Nakamura, S. Glucksberg, How about another piece of pie: the allusional pretense theory of discourse irony, J. Exp. Psychol. Gen. 124 (1995) 3.
[23] D. Sperber, D. Wilson, On verbal irony, Lingua 87 (1992) 53–76.
[24] A. Joshi, P. Bhattacharyya, M. Carman, Automatic sarcasm detection: a survey, 2016. arXiv Preprint arXiv:1602.03426 (accessed June 27, 2016) . http://arxiv.org/abs/1602.03426
[25] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: the figurative language of social media, Data Knowl. Eng. 74 (2012) 1–12.
[26] R. Giora, O. Fein, Irony: context and salience, Metaphor Symbol 4 (1999) 241–257.
[27] E. Sulis, D. Irazú Hernández Farías, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in Twitter: differences between #irony, #sarcasm and #not, Knowl. Based Syst. 108 (2016) 132–143, doi:10.1016/j.knosys.2016.05.035.
[28] B. Charalampakis, D. Spathis, E. Kouslis, K. Kermanidis, A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets, Eng. Appl. Artif. Intell. 51 (2016) 50–57, doi:10.1016/j.engappai.2016.01.007.
[29] O. Tsur, D. Davidov, A. Rappoport, ICWSM-a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews, ICWSM (2010) 162–169.
[30] R. Justo, T. Corcoran, S. Lukin, M. Walker, Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web, Knowl. Based 69 (2014) 124–133.
[31] F. Kunneman, C. Liebrecht, M. van Mulken, Signaling sarcasm: from hyperbole to hashtag, Inf. Process. Manage. 51 (2015) 500–509.
[32] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, p. 271.
[33] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 10, 2002, pp. 79–86.
[34] T. Zagibalov, K. Belyatskaya, J. Carroll, Comparable English-Russian book review corpora for sentiment analysis, in: Prooceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2010, pp. 67–72.
[35] C. Whissell, The dictionary of affect in language, in: R. Plutchnik, H. Kellerman (Eds.), Emotion: Theory, Research, and Experience, 1989, pp. 113–131.
[36] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, doi:10.1023/A:1010933404324.
[37] Y. Tausczik, J. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (2010) 24–54.
[38] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, LREC (2010) 2200–2204.
[39] E. Cambria, D. Olsher, D. Rajagopal, SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Pres, 2014, pp. 1515–1521.
[40] N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm, Mach. Learn. 2 (1988) 285–318, doi:10.1007/BF00116827.
[41] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[42] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27.
[43] N. Landwehr, M. Hall, E. Frank, Logistic model trees, Mach. Learn. 59 (2005) 161–205.
[44] S. Le Cessie, J. Van Houwelingen, Ridge estimators in logistic regression, Appl. Stat. (1992) 191–201.
[45] G. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.
[46] D. Rumelhart, G. Hinton, R. Williams, Learning internal representations by error propagation, in: Cognitive Science No. ICS-8506, California University San Diego La Jolla Institute, 1985.
[47] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[48] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC press, 1984.
[49] W. Cohen, Fast effective rule induction, in: Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 115–123.
[50] J. Pennebaker, R. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, Austin, 2015. doi:10.15781/T29G6Z.
[51] K. Kyle, S. Crossley, Automatically assessing lexical sophistication: indices, tools, findings, and application, Autom. Assess.Lexical Sophistication 49 (2015) 757–786.
[52] W. Parrott, Emotions in Social Psychology: Essential Readings, Psychology Press, 2001.
[53] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[54] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.
[55] P. Ravisankar, V. Ravi, I. Bose, Failure prediction of dotcom companies using neural network-genetic programming hybrids, Inf. Sci. 180 (2010) 1257–1267.
[56] A. Kulkarni, B.N. Kumar, V. Ravi, U.S. Murthy, Colon cancer prediction with genetics profiles using evolutionary techniques, Expert Syst. Appl. 38 (2011) 2752–2757.
[57] R.A. Johnson, D.A. Wichern, Applied Multivariate Statistical Analysis, 3rd ed., Prentice-Hall, 1992.
[58] M. Hall, Correlation-Based Feature Selection for Machine Learning, The University of Waikato, 1999.
[59] D.R. Cox, E.J. Snell, Analysis of Binary Data, 32, CRC Press, 1989.
[60] J. Payton, K. Weigandt, Satire and Data Science: An Exploration into One of the Current Final Frontiers, The City University of New York, 2015 https://rpubs.com/pm0kjp/satire_serious_reviews.
[61] I. Feinerer, K. Hornik, TM: text mining package, 2015. R package version 0.6-2 . http://cran.r-project.org/package=tm
[62] M. Porter, R. Boulton, Snowball stemmer, (2001).

[63] M. Bouchet-Valat, SnowballC: snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5.1., (n.d.).
[64] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. (2009) 11. http://www.cs.waikato.ac.nz/ml/weka.
[65] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (1971) 378–382.
[66] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[67] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1992) 1–58.
[68] K.N. Reddy, V. Ravi, Differential evolution trained kernel principal component WNN and kernel binary quantile regression: application to banking, Knowl. Based Syst.. 45–56 (2013).
[69] V. Srikrishna, R. Ghosh, V. Ravi, K. Deb, Elitist quantum-inspired differential evolution based wrapper for feature subset selection, in: A. Bikakis, X. Zheng (Eds.), International Workshop on Multi-Disciplinary Trends in Artificial Intelligence, Springer International Publishing, Fuzhou, China, 2015, pp. 113–124, doi:10.1007/978-3-319-26181-2_11.