Effective Analysis of Emotion-Based Satire Detection Model on Various Machine Learning Algorithms

Pyae Phyo Thu
University of Computer Studies, Mandalay
Mandalay, Myanmar
pyaephyothu149@gmail.com

Than Nwe Aung
University of Computer Studies, Mandalay
Mandalay, Myanmar
mdytna@gmail.com

***Abstract*—Even though the various features of satirical language have been studied in computational linguistics, most of the research works have relied on the performance of the single machine learning algorithm. However, the implicit traits embedded in the language demand more certain, precise and accurate combination powers of an individual algorithm. In this study, we analyzed the performance of emotion-based satire detection model on various machine learning algorithms: Regression, Naïve Bayes, SVM and ensemble classifiers. Experiments on shifting base classifiers to ensemble classifiers demonstrate that ambiguous and implicit nature of satirical emotions can lead to the misclassification accuracy while implementing the base classifiers but, offer reliable classification accuracy with ensemble classifiers.**

Keywords— Satire Detection; Emotional Features; Base Classifiers; Ensemble Classifiers

I. Introduction

Satire is a form of implicit communication which is used to expose human vices and follies in a humorous way. It can be widely found in everyday communication such as literature, television, the internet, social media, comics, and cartoons. For instance, "*A lot of people are afraid of heights. Not me, I'm afraid of width*", "*I talk to God but the sky is empty* - Sylvia Plath". Even though most of the implicit language detection have been focused on ironic language and sarcastic language, the study of satirical cues is just in a case of infancy.

The past works in implicit language detection have been emphasized on: (1) Lexical or specific textual features such as presence of adjectives, verbs, interjection [1, 7], (2) Contextual features which signal the contextual presuppositions between textual contents [6], and (3) Supervised and Semi-supervised Techniques which apply various classification algorithms in language detection [4, 10].

Recently in the area of machine learning, the concept of algorithms combination becomes one of the trending focuses to generate more precise and accurate system results. Moreover, analyzing satirical language with emotion is ambiguous [2], it sometimes leads to the misclassification accuracy. In this work, we investigate the effectiveness of the emotion-based satire detection model [8, and 9] by employing several machine learning algorithms.

In the rest of the paper, the problem of emotion-based satire detection is discussed in section II. And then, the detail of the satire detection model is explored in Section III. Later, experimental results and evaluation of the model is reported in section IV and then concluded it in section V.

II. Problem Definition

Existing works in satirical language detection mainly focused on lexical analysis. No prior work has been done on emotion-based satire analysis. However, some researcher [2] describe that satire can be characterized by emotions, but it is hard to detect due to their ironic dimension. Table I shows the frequency of each emotion involves in news article dataset. In order to analyze the emotions involved in the dataset, eight basic emotions extracted with the emotion lexicon: EmoLex are used.

TABLE I: FREQUENCY OF EMOTION IN NEWS ARTICLES DATASET

Emotion	**Frequency**	
	Satire	***Non-Satire***
Trust	1046	1153
Anticipation	268	273
Joy	42	92
Fear	220	118
Anger	61	31
Sadness	49	32
Surprise	11	4
Disgust	9	2

As shown in Table I, it can generally judge as positive emotions tend to be non-satirical language and negative emotions tend to be satirical language. However, if the rule negative emotions tend to be satire and positive emotions tend to be non-satire was applied, nearly half of the data will be misclassified. For instance, although 1,153 non-satire articles are correctly classified with emotion "Trust", about 1,046 satire articles are misclassified as non-satire.

Classifying on such ambiguous data with base classifiers can cause the misclassification. Hence, in this study, we demonstrate how the emotion-based satire model proposed in

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

[8, 9] differs in classification accuracy using different machine learning algorithms.

III. EMOTION-BASED SATIRE DETECTION MODEL

A corpus of 3,111 satirical and non-satirical news articles proposed by Frain and Wubben [5] was analyzed to distinguish between satirical and non-satirical language. Sentiment analysis and social cognition engine (SÉANCE) [3] was used to compute the emotional scores and sentiment scores. The details of the emotional features used in satire detection model [8, 9] are described as follow:

- **Emotion**: Eight basic emotional scores (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) extracted using SÉANCE's EmoLex are used as the basic emotional features.
- **Sentiment**: Positive, negative, neutral and compound sentiment scores extracted using SÉANCE's VADER lexicon are used as the basic sentimental features.
- **Bag-of Sorted Emotion (BOSE)**: With the hypothesis of intended emotion will differ in each satirical text, BOSE is ensemble from eight basic emotional scores and three sentimental scores as shown in Fig. 1. Afterward, the corresponding indexes of each sorted feature are extracted and concatenated to form BOSE.

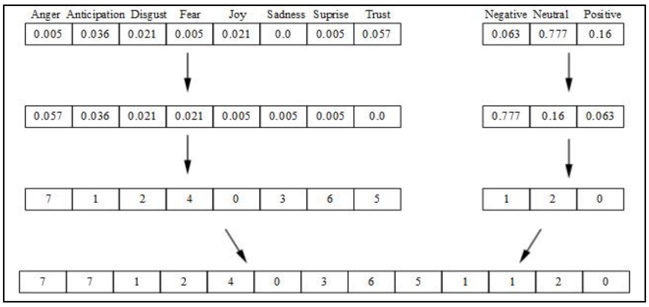


Fig. 1. Structure of Bag-of Sorted Emotion (BOSE)

- **SenticNet**: Four emotional dimension (sensitivity, aptitude, attention, pleasantness) and polarity scores computed by SÉANCE's SenticNet are used as supplementary emotional scores.
- **BOSE-TFIDF**: After occupying BOSE, emo-Bigram (e.g. Trust-Anticipation) and emo-Trigram (e.g. Trust-Anticipation-Disgust) are extracted to form emo-Bigram and emo-Trigram pairs. Later, term frequency-inverse document frequency (TF-IDF) weighting scheme is implemented on each EmoNgram to obtain the corresponding TFIDF scores.

IV. EXPERIMENTAL RESULTS AND EVALUATION

In this section, performance of the emotion-based satire detection model on various machine learning algorithms namely: Regression, SVM, Naïve Bayes, Bagging, AdaBoost, Random Forest, Extra Tree, and Gradient Boosting are reported. The classifiers are trained on 80% of balanced datasets (2,728 instances) and tested on 20% of the datasets (682 instances). Table II shows the F1 of the proposed features after employing each learning classifier.

The most reliable one in base classifiers is SVM which offers 0.682 F1 scores. Meanwhile, Regression gives the less performance score 0.551. In Ensemble classification, the best one is Random Forest (0.724), and the latter are Gradient Boosting (0.717), Extra Tree (0.710), Bagging (0.710) and AdaBoost (0.707) respectively. By analyzing the impact of the model in the more specific manner using confusion matrix, we found that the half of the satirical and non-satirical reviews are misclassified each other in base classification. This shows that base classifiers are confused in satirical language detection when they use such ambiguous features as emotions.

Fig. 2 demonstrates the error rate of the proposed model on each classifier. As shown in this figure, it's clear that all of the ensemble classifiers offer less misclassification rate than the base classifiers. Regression gives the highest error rate (nearly 0.45) among the classifiers. SVM is the best among the base classifiers which offers nearly 0.30 misclassification rate. Meanwhile, almost all of the ensemble classifiers misclassify about 25%.

Moreover, we compare the diagnostic ability of each classifier by varying its discrimination threshold as shown in Fig. 3. It shows that all classifiers are better than guessing. However, the best is the ensemble classifiers which offers more than 0.77 AUC score whereas the base classifiers give the less score ranging from 0.62 to 0.75.

TABLE II: AVERAGE F1 OF EACH FEATURE ON VARIOUS MACHINE LEARNING ALGORITHMS

Features	Base Classifiers			Ensemble Classifiers				
	Regression	*SVM*	*Naïve Bayes*	*Bagging*	*AdaBoost*	*Random Forest*	*ExtraTree*	*Gradient Boosting*
Emotion	0.585	0.652	0.622	0.672	0.654	0.682	0.682	0.676
Sentiment	0.525	0.576	0.607	0.657	0.655	0.655	0.654	0.658
BOSE	0.527	0.585	0.594	0.565	0.578	0.573	0.560	0.597
SenticNet	0.534	0.508	0.522	0.588	0.559	0.606	0.622	0.601
BOSE-TFIDF	0.472	0.573	0.575	0.547	0.573	0.553	0.568	0.565
ALL FEATURES	0.551	0.682	0.588	0.710	0.707	0.724	0.710	0.717

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

TABLE III: PERFORMANCE COMPARISON OF PROPOSED MODEL ON EACH CLASSIFIER USING CONFUSION MATRIX

Base Classifiers			**Ensemble Classifiers**				
Regression	**SVM**	**Naïve Bayes**	**Bagging**	**AdaBoost**	**Random Forest**	**ExtraTree**	**Gradient Boosting**
[215 136] [170 162]	[206 145] [72 260]	[172 179] [102 230]	[260 91] [107 225]	[252 99] [101 231]	[266 85] [103 229]	[257 94] [104 228]	[251 100] [93 239]

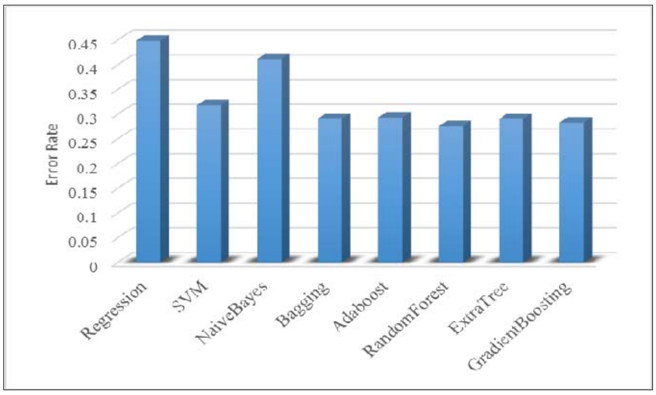


Fig. 2. Comparison of Error Rate on Various Learning Algorithms

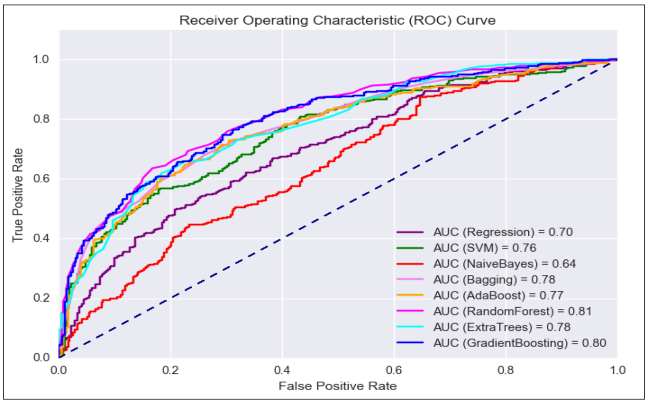
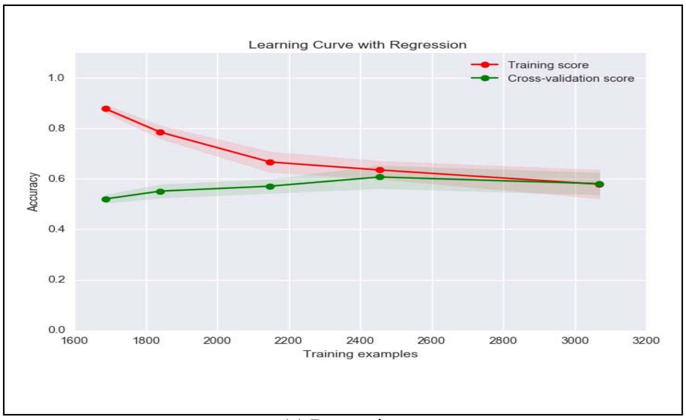


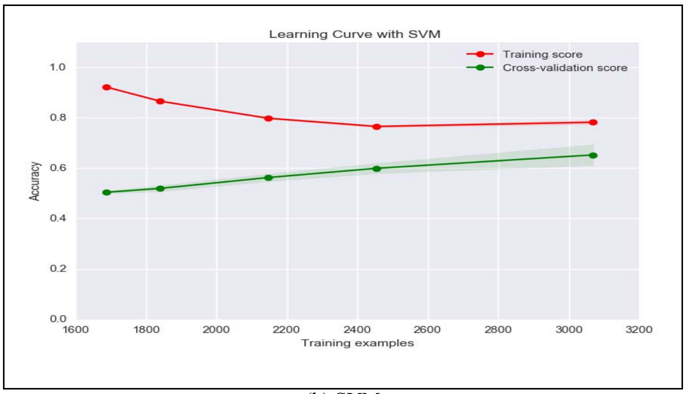
Fig. 3. Receiver Operating Characteristic (ROC) Curve on Various Learning Algorithms

Finally, we investigate the impact of each classifier by analyzing the learning curves achieved in each classification. Fig. 4 (a) shows the learning curves reagarding the evaluations of the base classifier: Regression; Fig. 4 (b) depicts the ones regarding the base classifier: SVM; Fig. 4 (c) illustrates the outcome of base classifier: Naïve Bayes; whereas Fig. 5 presents the five ensemble classifiers: Bagging, AdaBoost, Random Forest, Extra Tree and Gradient Boosting, respectively.

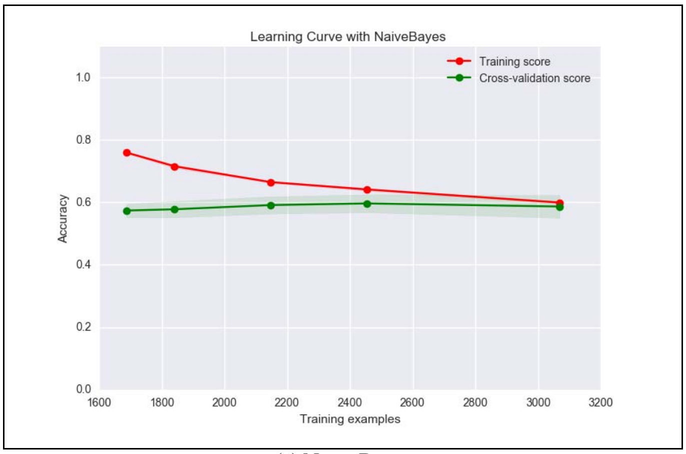
According to these figures, it is obvious that both of the performance on training score and 10-fold cross validation score using base classifiers do not perform well, i.e., greater the instances, lesser the training score. The highest cross-validation score that they offer is about 60%.



(a) Regression



(b) SVM

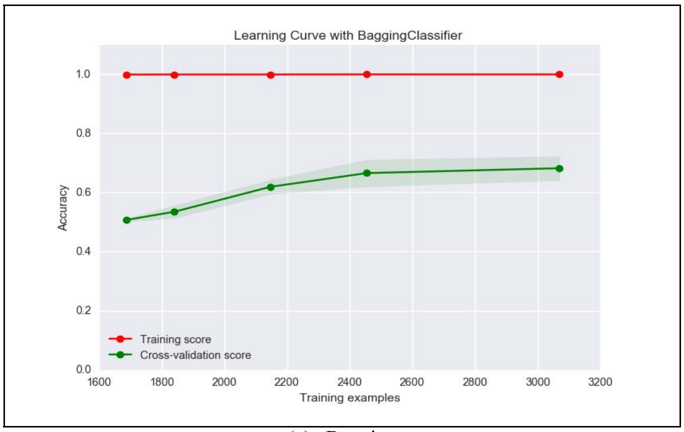


(c) Naïve Bayes

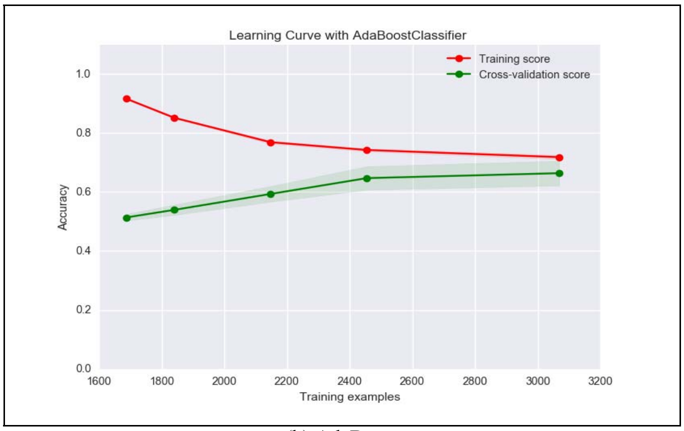
Fig. 4. Learning Curves regarding the Number of Training Examples with Base Classifiers

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

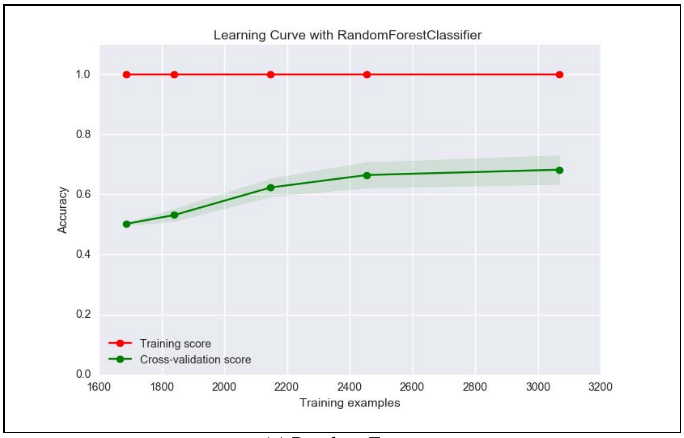
In ensemble classifications, both training score and cross-validation scores achieve better than in base classifications. Although they work with various training sizes, average cross-validation score gradually increase and offer about 70% accuracy. While training and testing the instances, most of the ensemble classifiers offer the highest score. However, AdaBoost and Gradient Boosting do not offer the highest training score as in other ensemble classifiers. The greater the data sizes, it slightly drops to 75% and 80% accuracy respectively.



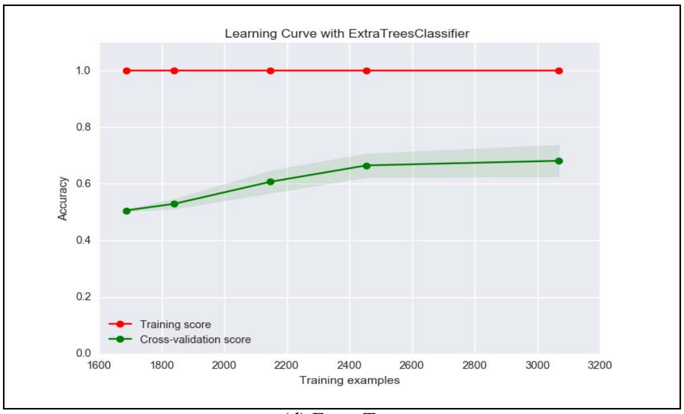
(a) Bagging



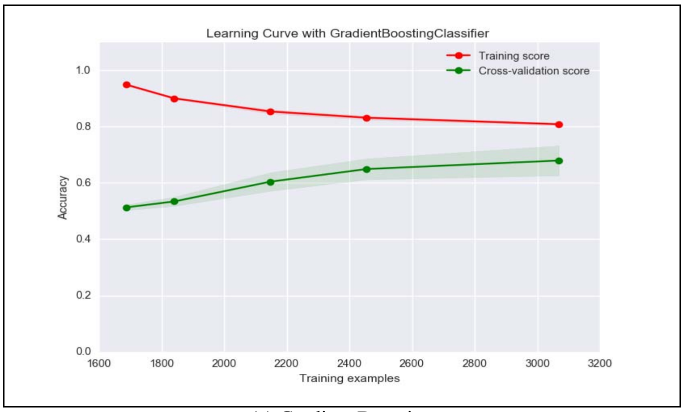
(b) AdaBoost



(c) Random Forest



(d) Extra Tree



(e) Gradient Boosting

Fig. 5. Learning Curves regarding the Number of Training Examples with Ensemble Classifiers

In conclusion, whether we compare the performance of base and ensemble classifiers using F1 score, Error Rate, ROC or Learning Curves, it's obvious that ensemble classifiers perform better than the base classifiers. Hence, ensemble classifiers are best suited for recognizing the indirect, ambiguous features embedded in the implicit language.

V. Conclusion

In this study, we analyzed the impact of ambiguous and implicit satirical features on various machine learning algorithms: Base classifiers and Ensemble classifiers. The results show that ensemble classifiers offer the better performance than the base classifiers in implicit language detection. However, due to the ambiguous nature of emotion and rareness of reliable emotion lexicon, the model only offer the reasonable classification accuracy (about 73%). In future, we plan to improve the performance of satire detection model using more reliable emotion lexicon.

References

[1] Burfoot C, Baldwin T. Automatic satire detection: Are you having a laugh?. InProceedings of the ACL-IJCNLP 2009 conference short papers 2009 Aug 4 (pp. 161-164). Association for Computational Linguistics.

[2] Barbieri F, Ronzano F, Saggion H. Do We Criticise (and Laugh) in the Same Way? Automatic Detection of Multi-Lingual Satirical News in Twitter. InIJCAI 2015 Jun 23 (pp. 1215-1221).

[3] Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behavior research methods. 2016 May 18:1-9.

[4] Fersini E, Pozzi FA, Messina E. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. InData Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on 2015 Oct 19 (pp. 1-8). IEEE.

[5] Frain, A. and Wubben, S., 2016. SatiricLR: a Language Resource of Satirical News Articles. In LREC.

[6] Khattri A, Joshi A, Bhattacharyya P, Carman MJ. Your Sentiment Precedes You: Using an author's historical tweets to predict sarcasm. In6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) 2015 Sep 17 (p. 25).

[7] Kreuz RJ, Caucci GM. Lexical influences on the perception of sarcasm. InProceedings of the Workshop on computational approaches to Figurative Language 2007 Apr 26 (pp. 1-4). Association for Computational Linguistics.

[8] Pyae Phyo Thu and Nwe Nwe, "Implementation of Satirical Emotion-Based Features on Satire Detection", 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2017). June 26-28, 2017, Kinrousha Plaza, Kanazawa, Ishikawa, Japan, pp. 149-154

[9] Pyae Phyo Thu and Nwe Nwe, "Impact Analysis of Emotion in Figurative Language", IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS 2017), May 24-26, 2017, Wuhan, China, pp. 209-214

[10] Tsur O, Davidov D, Rappoport A. ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. InICWSM 2010 May 23 (pp. 162-169).

978-1-5090-4045-2/17/$31.00 ©2017 IEEE