A Multilevel Approach to Sentiment Analysis of Figurative Language in Twitter

Braja Gopal Patra[1], Soumadeep Mazumdar[2], Dipankar Das[1], Paolo Rosso[3], and Sivaji Bandyopadhyay[1]

[1]Department of Computer Science & Engineering, Jadavpur University, Kolkata, India
[2]Microsoft India Development Center, Hyderabad, India
[3]PRHLT Research Center, Universitat Politècnica de València, Spain
{brajagopal.cse,mazumdar.soumadeep,dipankar.dipnil2005}@gmail.com
prosso@dsic.upv.es
sivaji_cse_ju@yahoo.com

Abstract. Commendable amount of work has been attempted in the field of Sentiment Analysis or Opinion Mining from natural language texts and Twitter texts. One of the main goals in such tasks is to assign polarities (positive or negative) to a piece of text. But, at the same time, one of the important as well as difficult issues is how to assign the degree of positivity or negativity to certain texts. The answer becomes more complex when we perform a similar task on figurative language texts collected from Twitter. Figurative language devices such as irony and sarcasm contain an intentional secondary or extended meaning hidden within the expressions. In this paper we present a novel approach to identify the degree of the sentiment (fine grained in an 11-point scale) for the figurative language texts. We used several semantic features such as sentiment and intensifiers as well as we introduced sentiment abruptness, which measures the variation of sentiment from positive to negative or vice versa. We trained our systems at multiple levels to achieve the maximum cosine similarity of 0.823 and minimum mean square error of 2.170.

Keywords: figurative text, sentiment analysis, sentiment abruptness measure, irony, sarcasm, metaphor

1 Introduction

With the rapid expansion of social media, a variety of user generated contents become available online. However, the major challenges are how to process the user generated contents such as texts, audio and images and how to organize them in some meaningful ways. It is observed that the existing systems achieved promising results for identifying opinions or sentiments along with polarities in case of literal language, because there is no secondary meaning embedded within it. In contrast, extracting sentiments from figurative language is one of the most challenging tasks in Natural Language Processing (NLP) because the literal meaning are discontinued and secondary or extended meanings are intentionally profiled. The affective polarity of the literal meaning may differ significantly from that of the intended figurative meaning [1]. Again, identifying the degree of the sentiment from these figurative texts are much more difficult. Figurative language contains several categories of tweets such as irony, sarcasm, and metaphor. The example below is a sarcastic tweet together with its degree of polarity in brackets.

you're such a cunt, I hope you're happy now #sarcasm (-4)

The study in Ghosh et al., [1] shows that metaphor, irony and sarcasm can each sculpt the sentiment of an utterance in complex ways, and texts limits the conventional techniques for the sentiment analysis of supposedly literal texts. For this reason, the analysis of sentiment degree in figurative language is considered to be a difficult tasks.

In this paper we present a novel approach for fine grained sentiment analysis of figurative language. Along with the features like parts of speech (POS), sentiment and intensifier, we also employed as further feature, "sentiment abruptness". We developed a multilevel classification framework to improve the performance of the system.

The rest of the paper is organized as follows: in Section 2, we describe the related work carried out in sentiment analysis of figurative languages. Section 3 describes the dataset. In Section 4, we describe the features and introduce the sentiment abruptness measure. The proposed multilevel system and its evaluations are described in Section 5. Finally, we concluded our study with future work in Section 6.

2 Related Work

Sentiment Analysis or Opinion Mining refers to the process of identifying the subjective responses or opinions about a specific topic. Much research have been conducted in the fields of opinion mining [18], sentiment extraction [10], emotion analysis [17] and review sentiment analysis [4]. Recent publications in the field of sentiment analysis were based on user generated data collected from different social media platforms like Facebook, Twitter, reviews and blogs etc. [2], [4], [19].

Extracting polarity (positive or negative) from the text is one of the tasks in sentiment analysis and used to achieve high accuracies [1]. Unfortunately, sentiment analysis when figurative language is employed still remains a challenging research topic as the languages have secondary or extended meanings and are intentionally profiled [2]. There have been several automatic computational approaches were attempted to categorize the figurative texts, such as humor recognition, metaphor or irony or sarcasm detection [2], [3].

Limited amount of research has focused on identifying the degree of sentiment in the figurative language. A shared task on "*Sentiment Analysis of Figurative Language in Twitter*" was organized in SemEval-2015 [1]. One of the main goals of this task was to evaluate the degree to which a conventional sentiment analysis approach suits for creative language or figurative language. In the task, a set of tweets that are rich in irony, metaphor, and sarcasm were given and the goal was to determine whether the user has expressed a positive, negative or neutral sentiment in each, and the degree to which this sentiment has been communicated. To capture the degree or intensity of the irony, metaphor and sarcasm, each of the participating systems were asked to assign fine grained sentiment score in a scale of -5 to +5.

A total of 15 teams participated in the above task and the team *CLaC* achieved the top results among all other submitted systems [1]. Features like unigram, bigram, parts-of-speech (POS), and sentiment lexicons such as SentiWordNet [14], NRC Emotion lexicon [13], and AFINN dictionary [11] were used in most of the systems [5], [20]. Support vector machines (SVMs), LibSVM (a variant of SVMs), and Decision Trees are the main classifiers that were used to develop several sentiment analysis systems for figurative language [1], [5], [20].

3 Dataset

We used the same datasets (trial, training and test) as provided by the organizing committee of the SemEval-2015 Shared Task-11[1] for our experimentation purpose. The trial and training datasets consist of 906 and 8000 tweets where each of the instances is accompanied with a real valued score ranging from [-5, +5]. Similarly, the test dataset contains 4000 tweets accompanied with absolute valued score from [-5, +5].

The distributions of each tweet class with respect to the trial, training and test datasets are shown in Table 1. We have also counted the frequencies of well-established hashtags like #sarcasm, #irony and #not and recorded the statistics in Table 1. We found a total of 4077, 257 and 1242 unique hashtags in training, trial and test datasets, respectively. We removed the junk characters from the tweets. We also normalized the words having multiple characters (for example, the word ‘yesssss’ to ‘yes’ using an English dictionary).

Table 1. Tweet class distribution and hashtag statistics.

Data Set	**Tweet Score**											**Total Tweets**	**Hashtags**		
	-5	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**	**5**		**#irony**	**#not**	**#sarcasm**
Trial	5	80	359	227	75	46	46	33	27	7	1	906	720	472	455
Train	6	364	2971	2934	861	345	165	197	106	49	2	8000	1405	3328	1975
Test	4	100	737	1541	680	298	169	155	201	111	4	4000	32	45	197

[1] http://alt.qcri.org/semeval2015/task11/

4 Feature Analysis

In order to develop an automatic tweet classification system based on the above datasets, we identified the basic textual and semantic features as available in the literature [5]. We have considered the following key features like *POS*, *sentiment features*, *intensifiers* and *sentiment abruptness* measure for tweet polarity strength classification tasks.

4.1 POS (I)

POS tag plays an important role in sentiment analysis [4]. Thus, we have used the ark-tweet-nlp [12] tool to parse each of the tweets to find out the POS tags of each word and included them as a feature in our experiments. The POS feature is used only for the Conditional Random Field (CRF) [8] based system.

4.2 POS Sequence (II)

We have observed that the POS sequence also plays important role in sentiment analysis [4]. For e.g., '*brave/JJ heart/NN*', here the word 'brave' is an adjective and it enhances the positivity of the noun 'heart'. Therefore, we utilized this information while training our datasets using CRF.

4.3 Intensifier (III)

We prepared six types intensifier lists for our system. Generally, such words help in identifying the intensity of the sentiments. Basically, an intensifier emphasizes or reduces the sentiment value or effect of the sentiment word it is preceded or followed by. For example, if there are two sentences, 'I am sad.' and 'I am very sad', in the second sentence, the word 'very' emphasizes the degree of negativity of the word 'sad' in the sentence. The intensifier classes and examples from each class are given in the Table 2.

We assigned values for each of the intensifier classes, for e.g., we assigned 2, 1.5, 1, -1.5, and -2 values for maximizers, boosters, approximators, compromisers, diminishers, and minimizers respectively. If an intensifier is found in a tweet before a sentiment word, its corresponding value is multiplied with the succeeding positive or negative value of the sentiment word. The positive or negative value of the sentiment word is identified using SentiWordNet.

4.4 Sentiment Feature (IV)

Sentiment lexicons are the most important features for any kind of sentiment identification or classification tasks [4, 5]. We have used several lexicons like SentiWordNet [14], WordNet Affect [6], SentiSense Synset [9], Effect WordNet [15], AFINN dictionary, NRC Word-Emotion Association Lexicon [13], Taboada adjective list [10] and Whissell dictionary [16] to identify the sentiment/emotion

Table 2. List of intensifiers.

Intensifier class (total instances)	**Words**
Maximizers (14)	completely, absolutely, totally, thoroughly, etc.
Boosters (38)	very, highly, immensely, exceedingly, etc.
Approximators (12)	nearly, virtually, effectively, all but, etc.
Compromisers (5)	fairly, pretty, rather, etc.
Diminishers (17)	slightly, a little, a bit, somewhat, rather, moderately, etc.
Minimizers (8)	hardly, scarcely, barely, almost, etc.

class of the words and used as features. For the CRF based system, each of the words in a tweet is marked with either positive or negative or neutral using all of the above lexicons. Whereas for the other system, we counted the total number of positive and negative words present in a tweet using the above lexicons.

4.5 Sentiment Abruptness (V)

Finally, in the present work, we proposed a special measure named as *Sentiment Abruptness*, which measures the variation of sentiment from positive to negative or vice versa in a tweet text. We plot each of the sentiment tokens on a graph with the help of SentiWordNet scores on the Y-axis and the token position of a tweet on the X-axis. Let us consider two sample tweets, T_1 and T_2.

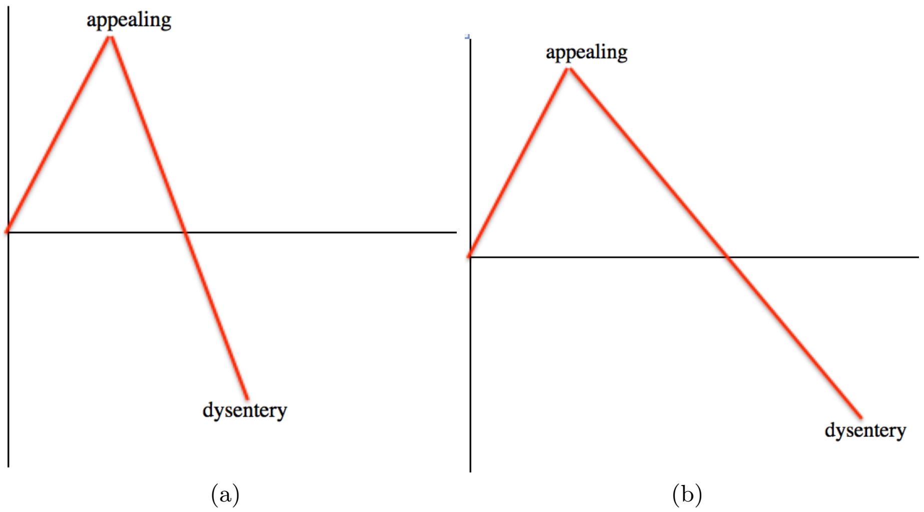
***T_1**: RT @TheeJesseHelton: "A Million Ways To Die In The West" looks about as **appealing** as **dysentery**. (-4).*

***T_2**: RT @TheeJesseHelton: "A Million Ways To Die In The West" looks like an **appealing** movie which I missed because of **dysentery**. (0).*

If we consider '*appealing*' and '*dysentery*' as the only sentiment points on the above tweets, the two sentiment plots would look somewhat as shown Figure 1.

Two sentiment words, the positive and negative sentiment values extracted from SentiWordNet in both of the tweets are same. But, we get different types of sentiment curves because of the difference in the vicinity of the two words. The "sharp turns" in the sentiment curve indicates higher level of sarcasm in the tweet.

Consider the following sentiment plot for a tweet in Figure 2. Thus, the 'turn' or the degree of polarity can be detected by measuring the curvature of a circle passing through a given triplet of points (P_1 (x_1, $y_{1)}$, P_2 (x_2, y_2), and P_3 (x_3, y_3)) in the sentiment curve (as a higher curvature indicates a sharper turn). We calculate the afore-mentioned curvature using the well-known Menger's curvature formula [7] indicated in equation 1. Thus we propose to detect the degree of the sentiment with the help of Menger's curvature and using the coordinates of the points, the abruptness score (K) is given by:



(a)

(b)

Fig. 1. Sentiment plots of T_1 (a) and T_2 (b)

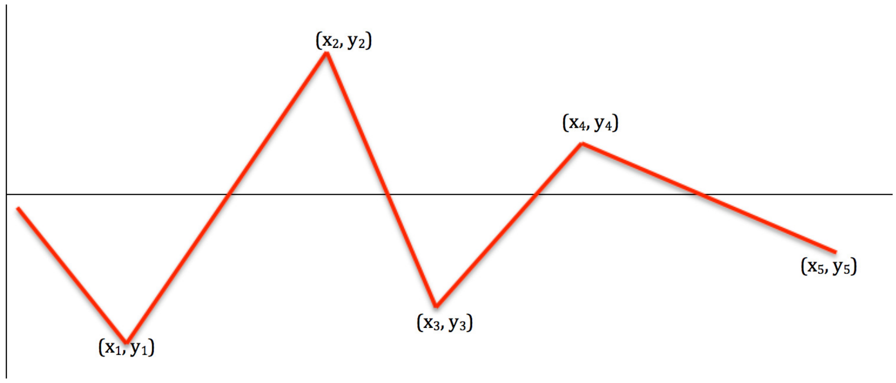


Fig. 2. Sample sentiment plot with five sentiment points.

$$\mathbf{K} = \frac{4 \times f(P_1, P_2, P_3)}{\sqrt{g(P_1, P_2, P_3)}} \tag{1}$$

Where the area of triangle using the coordinates,

$$\mathbf{f(P_1, P_2, P_3)} = \frac{1}{2} \left|(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)\right| \tag{2}$$

The product of the three sides is $\sqrt{g(P_1, P_2, P_3)}$ and

$$\mathbf{g(P_1, P_2, P_3)} = ((x_2-x_1)^2+(y_2-y_1)^2)((x_3-x_1)^2+(y_3-y_1)^2)((x_3-x_2)^2+(y_3-y_2)^2) \tag{3}$$

However, we have to normalize the sentiment abruptness score of a curve since a tweet with multiple sentiment points would score more than a tweet with a few sentiment points but with similar intensity. Therefore, we just divide the total score by the number of tokens present in the tweet. The algorithm used for calculating the sentiment abruptness score is given below.

Algorithm (Sentiment abruptness):

```
1: Initialize $Total_{abruptness} = 0$;
2: for each tokens in tweet:
    a. if token in SentiWordNet then put it in SList
3: for each triplet of points $P_1$, $P_2$, $P_3$ in SList:
    a. Calculate sentiment abruptness score (K);
    b. $Total_{abruptness} = Total_{abruptness} + K$;
4: $Measured_{abruptness} = Total_{abruptness}$ / SList_length;
5: return $Measured_{abruptness}$;
```

We calculated the average sentiment abruptness score for each of the tweet classes. The values for the training dataset are as follows, 0.132039, 0.127876, 0.111354, 0.095409, 0.09699, 0.090894, 0.093322, 0.101701, 0.109143, 0.126745 and 0.266667 for -5, -4, -3, -2, -1, 0, +1, +2, +3, +4, and +5 tweet scores, respectively. We can observe that the sentiment abruptness score is higher for -5 and +5, whereas lower for the 0 tweet score. The sentiment abruptness score for each of the tweets is calculated using the above algorithm and this score is used as a feature for our experiments.

5 Multilevel Training and System Framework

The trial and training dataset are annotated with real valued scores whereas the test dataset is annotated with absolute values. Thus, we developed two basic systems; (a) a regression model followed by classification to cope up with real valued data and (b) a classification model capable to work on the absolute valued scores.

5.1 Evaluation Criteria

The performance of our model was evaluated based on the cosine similarity (CS) of the desired output with the system output as proposed in the Task 11 of SemEval-2015 [1]. We also evaluated performance of the systems using Mean Squared Error (MSE) metric. The equations for the CS and MSE are given in the equation 4 and 5 respectively.

$$\mathbf{CS} = \frac{Actual.Predicted}{|Actual|\,|Predicted|} \tag{4}$$

$$\mathbf{MSE} = \frac{\sum_{i=1}^{n}(Actual_i - Predicted_i)^2}{n} \tag{5}$$

5.2 System 1

We developed our first system using the real valued scores of training and trial dataset. The test dataset contains one absolute scores, thus we have not used this for the first system. A total of 8906 tweets have been used to develop the first system. We used the CRF model first and then the Support Vector Machine Regression model of Weka[2] to build a multilevel classification framework.

Initially, to build the CRF model, we rounded off the scores of training and trail dataset to absolute values. We performed the 10-fold cross validation and in the first level, CRF classifier gives the maximum CS of 0.746 and minimum MSE of 3.096. The performance of the system according to each of the features and their combinations are provided in Table 3.

Table 3. Performance of the CRF based classifier in System 1.

Features	**CS**	**MSE**
I	0.736	3.270
I+ II	0.739	3.220
I + II + III	0.741	3.201
I + II + III + IV	**0.746**	**3.096**

The output of CRF, i.e. the absolute score of each tweet is then used as a feature in the second level regression model. We achieved the maximum CS of 0.823 and minimum MSE of 2.170 using the 10-fold cross validation. It was observed that CS did not vary even with the inclusion of CRF output as a feature, but the MSE reduced significantly. The CS and MSE with respect to the various features are given in Table 4.

[2] http://www.cs.waikato.ac.nz/ml/weka/

5.3 System 2

We rounded off the scores from real values to integer for both trial and training datasets in order to consider the problem as a classification problem rather than regression. Therefore, we trained our second system using the absolute valued score of each tweet of trial and training dataset. We rounded off all the scores of tweets in the training and trial dataset and merged them together, i.e. we have a total of 8906 tweets for training. We used LibSVM classifier of Weka for classification purpose and tested the system on the test dataset of 4000 tweets.

Our second system achieves the maximum CS of 0.765 and minimum MSE of 2.973 as shown in Table 4. The performance obtained by this second system shows a marginal improvement of CS (0.007) over the best performing system of Task 11 at SemEval-2015 [5].

Table 4. Performance of the proposed systems.

Features	**System 1**		**System 2**	
	CS	**MSE**	**CS**	**MSE**
III + IV	0.802	2.227	0.563	4.061
V	**0.737**	**3.265**	**0.656**	**3.357**
III + IV + V	0.822	2.230	**0.765**	**2.973**
III + IV + V + CRF Class	**0.823**	**2.170**	X	X

The noticeable improvement was found when we incorporated sentiment abruptness as a feature. This feature solely gives the maximum CS of 0.737 and 0.656 and minimum MSE of 3.265 and 3.357 for system 1 and system 2 respectively. Therefore, we added this measure into our existing feature set and the corresponding results are found in Table 4. One of the major problems faced during the experiment was to handle unequal number of tweet instances present in the datasets. The total count of the instances having scores -2 and -3 was larger than others and we observed the biasness of the classifiers towards these two classes.

6 Conclusion and Future work

In this paper we introduced a new measure, sentiment abruptness which achieves the maximum CS of 0.737 and 0.656 for the respective systems for identifying sentiment scores of the figurative texts. The system achieves the maximum CS of 0.823 with the help of multilevel classification along with other features.

In future, we plan to use the tweet dependency parsers to get the relations for different phrases. If the relations in a tweet are contradictory, then there may be a chance of tweet to be ironic or sarcastic. Another immediate goal is to develop some lexicons or ontology from the tweet data for sentiment analysis as developed in [5] and use this ontology to detect figurative language in social media texts. Moreover, we plan to consider different sentiment lexicons considering the abruptness measure.

Acknowledgments. The work reported in this paper is supported by a grant from the project "CLIA System Phase II" funded by Department of Electronics and Information Technology (DeitY), Ministry of Communications and Information Technology (MCIT), Government of India. The work of the fourth author is also supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAPATER (PrometeoII/2014/030).

References

1. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Reyes, A., Barnden, J.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: 9th International Workshop on Semantic Evaluation (SemEval), Co-located with NAACL, Denver, Colorado, pp. 470–478. Association for Computational Linguistics (2015)
2. Reyes, A., Rosso, P. Veale, T.: A multidimensional approach for detecting irony in twitter. Language Resources and Evaluation 47(1), 239–268 (2013)
3. Reyes, A., Rosso, P., Buscaldi, D.: From humor recognition to irony detection: The figurative language of social media. Data & Knowledge Engineering 74, 1–12 (2012)
4. Patra, B.G., Mandal, S., Das, D., Bandyopadhyay, S.: JU_CSE: A conditional random field (CRF) based approach to aspect based sentiment analysis. In: 8th International Workshop on Semantic Evaluation (SemEval), Co-located with COLING, Dublin, Ireland, pp. 370–374. Association for Computational Linguistics (2014)
5. Ozdemir, C., Bergler, S.: CLaC-SentiPipe: SemEval2015 subtasks 10 B, E, and task 11. In: 9th International Workshop on Semantic Evaluation (SemEval), Co-located with NAACL, Denver, Colorado, pp. 479–485. Association for Computational Linguistics (2015)
6. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: 4th International Conference on Language Resources and Evaluation, pp. 1083–1086 (2004)
7. Léger, J.C.: Menger curvature and rectifiability. Annals of mathematics 149, 831–869 (1999)
8. Lafferty, J. D., McCallum, A., Pereira, F. C. N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: 18th International Conference on Machine Learning, pp. 282–289 (2001)
9. Albornoz, J. C. de, Plaza, L., Gervás, P.: SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In: 8th International Conference on Language Resources and Evaluation, pp. 3562–3567 (2012)
10. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics 37(2), 267–307 (2011)
11. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A. C.: Bad news travel fast: A content-based analysis of interestingness on twitter. In: 3rd International Web Science Conference, ACM (2011)
12. Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N. A.: Improved Part-of-Speech tagging for online conversational text with word clusters. In:NAACL. Association for Computational Linguistics (2013)

13. Mohammad, S., Turney, P.: Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence 29 (3), 436–465 (2013)
14. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: 7th conference on International Language Resources and Evaluation, Valletta, Malta (2010)
15. Choi, Y., Wiebe, J.: +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In: EMNLP (2014)
16. Whissell, C., Fournier, M., Pelland, R., Weir, D., Makarec, K.: A dictionary of affect in language: IV. Reliability, validity, and applications. Perceptual and Motor Skills, 62(3), 875–888 (1986)
17. Patra, B.G., Takamura, H., Das, D., Okumura, M., Bandyopadhyay, S.: Construction of Emotional Lexicon Using Potts Model. In: International Joint Conference on Natural Language Processing (IJCNLP), pp. 674–679 (2013)
18. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2, 1–135 (2008)
19. Vilares, D., Alonso, M. A., Gómez-Rodríguez, C.: On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. Journal of the Association for Information Science and Technology 66(9) 1799–1816 (2015)
20. Barbieri, F., Ronzano, F., Saggion, H.: UPF-taln: SemEval 2015 Tasks 10 and 11 Sentiment Analysis of Literal and Figurative Language in Twitter. SemEval-2015, 704–708 (2015)