A New Framework for Sentiment Analysis with Six-Tuples

Debanjan Banerjee
Dept. of MS-SDE
Manipal Universal Learning
Kolkata - 700091, India

Bikromadittya Mondal
Dept. of Computer Science & Engineering
B P Poddar Institute of Management & Technology
Kolkata - 700052, India

Sarit Chakraborty
Dept. of Computer Science & Engineering
B P Poddar Institute of Management & Technology
Kolkata - 700052, India

ABSTRACT

Text analytics is one of the growing fields of interest from the scientific and the business communities in recent times.In this paper a new framework is introduced which consists of six elements for presenting a structured and organized form of describing any opinionated sentiment. The basic elements of this frameworkare opinion holder, an entity which is the intended target of the opinion, time of the expressed opinion, sentiment of the opinion, aspect or attribute of the opinion and representation of the opinion. This framework has been constructed keeping in mind the importance of the influence an opinion can inflict in the minds of those to whom the opinion is expressed by the opinion maker. The framework has been assumed from the perspective of the potential influence an opinion can have on the receivers of the opinion. The proposed framework is more improved than other existing works done so far.

Keywords

Sentiment analysis framework, Opinion mining, Text analytics, Natural language processing.

1. INTRODUCTION

One of the prime objectives for the study of text analytics attempts to understand these patterns in day to day communications. One of the most observable trends of social network is that people manages to influence others with their ideas, judgments and opinions. Understanding influence through the expression of online subjective comments is an important problem in the realm of text analytics. It is important to analyse potential factors which play an important role behind how any potential idea would gain influence in a social sphere. This paperintroduces this particular framework keeping in mind the potential influence the opinion can have from those who experience this particular opinion. This workassumes that an opinionis represented either egoistically or intelligently. By egoistic representation of the opinion the work understands that the opinion which is expressed based upon a particular person's bias, personal belief, likings or feelings without expressing the clear logic or reason in support of his viewpoint. So for example when an opinion maker states that "I feel the book is bad" here the assumption is that the person is stating his egoistical feeling about the subject without providing the necessary reason for his emotions. Whereas in the intelligent representation of opinion consists of those opinions whereby the holder of the opinion includes a particular reason for his liking or disliking of a particular subject. For example, when the person opines that "I believe the book is bad because the prose is boring", we recommend classifying this particular opinionated expression as intelligent representation since this opinionated representation includes a very specific reason part behind his stated emotion towards the subject. Now the intelligent representation can be explicit when the person uses words "because", "so", "therefore" to specifically distinguish the reason part from that of the opinion part whereas there are explicit intelligent representations whereby the reason or logic behind the person's opinion tends to be implicit for example "you can invest in the car ; it's mileage is great." Here though the opinionated expression is explicit intelligent representation since no specific word such as because, since, therefore has been used in this context. The work observes that both egoistically as well as intelligent representation of the human opinion can be influential in specific circumstances. The framework has been deliberated with sentence level data.

2. RELATED WORKS

The research community has been working on studying different aspects of social media for the past few years. Wiebe, Bruce and O'Hara [4] came up with the idea of classifying the opinionated expressions into positive, negative or neutral classes. Liu [1], [2], [3] proposed the organization of an opinion into five specific parts which are the target of the opinion, the aspect of the target on which the opinion has been expressed, opinion holder, the time in which the opinion has been expressed and opinion sentiment. Our work is different from these works since these works primarily consider opinion from the point of view of the opinionated expression whereas our framework is based upon how an opinion is expressed from the viewpoint of the opinion's potential influence towards the particular work. This is why we propose the six tuple framework whereby we improve upon Bing Liu's work by expanding the definition of the opinion to include the way the opinion has been represented to the outside world for understanding the opinion's potential influence. Arjun Chaudhury [5] came up with the concept of the rationally evaluated opinion and emotionally evaluated opinion whereby he argues that human beings evaluate an opinion in two ways. The emotionally evaluated opinion is expressed based upon the feelings of the opinion holder and the rationally evaluated opinions are based upon their real world utilitarian experiences with that particular product. Our work is distinct from this work that our framework considers the opinion from the opinion's potential influence of the same instead of the emotional motivation behind the opinion.

Gain ratio has been a very prominent technology for feature selection for text classification. Morariu et al. [6] used gain ratio for improving accuracy and they achieved more than 90 percent accuracy with only less than 25 percent features. Zia et al [7] used naïve bayes alongside gain ratio for feature selection for urdu text categorization. Phayung et al. [8] used chi-square technique and got more than ninety percent accuracy in text classification. Debole and Sebastiani [10] used supervised term auditing for automated text categorization. Ikonomakis et al. [11] used a cumulative technique for text classification. Karegowda et al. [12] used correlation and gain ratio for feature selection. Nicolosi emphasizes the usage of gain ratio for text classification [13]. Singh et al. [14] used gini coefficient for online text classification. Shang et al.[15] used a novel technique called maximizing global information gain for text classification. Yang and Pedersen [16] with their comparative study on text categorization outlined gain ratio as a prominent feature selection technique.

Novovicova [17] et al., used improved mutual information for text categorization feature selection. Rogati and Yang [18] used naïve bayes alongside information gain feature discrimination technique for high-performance text classification. Patil and Mohammad [19] used rough set theory for improving accuracy for text categorization. Zhaouhui et al. [20] used gain ratio for imbalanced data.

3. PROPOSED WORK

Our framework considers an opinion to be made of six elements which can be represented in the format of (*ei*, *aij*, *sijkl*, *hk*, *tl* ,*ri*) In the framework , *ei* is an entity which is the main target of the opinion , *aij* is the particular attribute of the entity upon which the opinion has been expressed , *sijkl* is the particular sentiment which has been expressed about the particular entity and the sentiment is positive or negative about the particular entity, *hk* is the holder of the particular opinion, *tl* is the particular time when the opinion has been expressed and *,ri* represents the way opinion holder chooses to represent the opinion to others. We assume that the first five elements of an opinionated expression are known to us whereas our primary interest lies in classifying the six-th element of the framework which happens to be the classification of the opinionated expression into two classes i.e. emotional and intelligent.

3.1 Algorithm for Classifying the Opinions

Algorithm for deriving the representation of the opinion considers that the intelligent opinions have two distinct parts which are a personal expression part (I detest this coffee) and a logic or reason part (if it tastes so bitter). Whereas the emotional opinions only consist of the personal expression part i.e. (I love this coffee.) which only represents the personal expression part.So the algorithm for deriving opinion representation classes

(1) For any given opinionated sentence, deconstruct it into a decision tree structure whereby the root node of the tree is the personal expression part of the opinion and the reason part(s) should be considered as the child nodes.

(2) If the tree structure has only the root node then consider the opinionated statement as an emotional opinion.

(3) If the tree structure has both the root node and the child node(s) then consider the opinionated statement as an intelligent opinion.

Let us consider two opinionated expressions i.e. “I love this book.” And “I love this book because the story is good”. The first opinionated expression is: “I love this book”. Now since in this case there is no specific reason given behind the opinionated expression according to algorithm, there is only emotional part involved and we are unable to form any child node and thereby we classify the representation the opinion as of “emotional”. In the second opinionated expression we find that there is both a reason part as well as an emotional part expressed by the opinion maker and thus the representation behind this particular opinion can be expressed as “intelligent”.

Statement 1: Even worse than The Bell Curve.

Statement 2: This was an incredibly frustrating book.

Statement 3: Granted, they have their own share of economic problems, but the larger industrial powers aren't as bad off as we are.

Statement 4: Murray looks at only part of what happened in America during the half-century covered in the book, so his analysis and recommendations seem deeply blinkered.

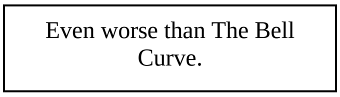
Statement 5: The bad news: he thinks this happened--in part--because the working class got indolent.

Statement 6: It is a myopic view of history, claiming that the social policies of the Great Society in the 1960s were responsible for the moral decay of the American public.

Now after applying the above algorithm to the below statements, we get the following outcomes.

Statement 1: Even worse than The Bell Curve.

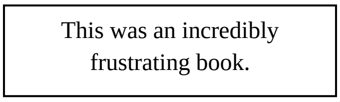
The Deconstructed tree:



Classification outcome: Since there is only root node present in the deconstructed tree, the opinion representation should be classified as egoistic.

Statement 2: This was an incredibly frustrating book.

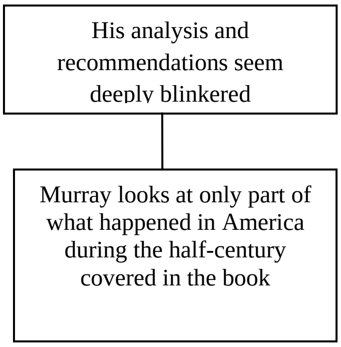
The Deconstructed tree:



Classification outcome: Since there is only root node present in the deconstructed tree, the opinion representation should be classified as egoistic.

Statement 3: Murray looks at only part of what happened in America during the half-century covered in the book, so his analysis and recommendations seem deeply blinkered.

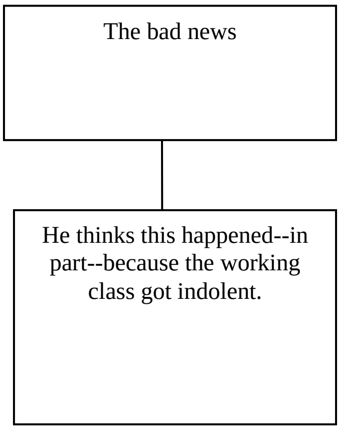
The Deconstructed tree:


His analysis and recommendations seem deeply blinkered
Murray looks at only part of what happened in America during the half-century covered in the book

Classification outcome: Since there is a root node and a child node present in the deconstructed tree, the opinion representation should be classified as intelligent.

Statement 5: The bad news: he thinks this happened--in part--because the working class got indolent.

The Deconstructed tree:


The bad news
He thinks this happened--in part--because the working class got indolent.

Classification outcome: Since there is a root node and a child node present in the deconstructed tree, the opinion representation should be classified as intelligent.

4. RESULTS

We have implemented the proposed methods and tested on data set obtained from goodreads.com book review website for the book “Coming apart” by Charles Murray. For the framework there was a human inter-annotator agreement whereby we decided to settle that a particular set of keywords and presence/absence of these keywords. This framework restricts our work to specifically those opinionated expressions which are explicitly intelligent. These keywords are

Keywords	Because , so , and

Our methodology employs the usage of machine learning algorithms i.e. the Random Forest and the conditional probability based algorithm Naïve Bayes. We concentrate on using keywords because, so, and as features for the machine learning algorithm. While pre-processing of data we tend to include those sentences as our training and test data which includes the keywords because, so, and, thus we are able to use these sentences also as part of our training data. We use Weka 3.6.9 data mining tool for our data analysis and results verification and we tried decision-stomp classifier and traditional naïve Bayes classifier for our purpose.

4.1 Feature Discriminators

For feature discrimination the work utilizes two well-renowned feature discriminating items such as Information gain and gain ratio.

4.1.1 Information gain

The information gain attribute can be expressed as

$$IG \leftarrow 1 - Entr.$$

Whereas $Entr.$ represents entropy for each individual word.On the other hand Entropy can be represented as

$$Entr. \leftarrow P(W) * \log P(W)$$

Where

$$P(W) = Probability\ of\ the\ indivisual\ word$$

Probability of the individual word $P(W)$can be derived by the following ratio

$$P(W) = \frac{T}{A}$$

Where

$$T = Total\ number\ of\ occurrences\ of\ the$$

$$individual\ wordin\ the\ corpus$$

A= $Total\ number\ of\ occurrences\ of$

$$all\ the\ wordsin\ the\ corpus$$

The work proceeds to prepare rankings based upon the information gain for all the attributes (the value is usually less than 1) to use this measure in classification techniques. The classification techniques which use this attribute are J48, random forest, K-nearest neighbour, naive bayes and SVM, respectively.

Let $Attr$ be the set of all attributes and $Ex$the set of all training examples, $value\ (x, a)$with $x \in Ex$ defines the value of a specific example $x$for attribute $a \in Attr$,Hspecifies the entropy. The information gain for an attribute $a \in Attr$is defined as follows:

$$IG\ (Ex, a) = H(Ex) - \sum_{v \in values\ (a)} \frac{|\{x \in Ex\ |value(x,a) = v\}|}{|Ex|} \cdot H\ (\{x \in Ex\ |value(x,a) = v\})$$

Table 1.Experimental results

Keyword	Classifier	Training (%)	Testing (%)	Feature discriminator used	Accuracy (%)
And	Random Forest	60	40	Information Gain	70
And	Naïve Bayes	60	40	Information Gain	71
And	Random Forest	50	50	Gain ratio	68
And	Naïve Bayes	50	50	Gain ratio	70
Because	Random Forest	60	40	Information Gain	56
Because	Naïve Bayes	60	40	Information Gain	11
Because	Random Forest	50	50	Gain ratio	28
Because	Naïve Bayes	50	50	Gain ratio	33
So	Random Forest	60	40	Information Gain	82
So	Naïve Bayes	60	40	Information Gain	50
So	Random Forest	50	50	Gain ratio	81
So	Naïve Bayes	50	50	Gain ratio	25

4.1.2 Gain ratio attribute

The In this technique we evaluate the worth of an attribute by measuring the gain ratio with respect to the class. By adopting this technique we are able to discriminate against those attributes which have large numbers of distinct values. This fact benefits the usage of gain ratio for many decision-tree based classifiers for improving their accuracies. The attributes with the highest gain ratio are considered most favourable to be used in classification techniques such as the J48, random forest, K-nearest neighbour, naive bayes and SVM, respectively.

The intrinsic value for a test is defined as follows:

$$IV\ (Ex, a) = -\sum_{v \in values\ (a)} \frac{|\{x\ \in Ex\ |value(x,a) = v\}|}{|Ex|} * log_2 \left(\frac{|\{x\ \in Ex\ |value(x,a) = v\}|}{|Ex|} \right)$$

The information gain ratio is just the ratio between the information gain and the intrinsic value:

$$IGR\ (Ex, a) = IG/IV$$

From our experiments we observe that the performance of the Naïve Bayes classifier improves as the percentage of training data is enhanced with respect to the testing data. The performance of the Random Forest algorithm also improves constantly as we tend to enhance the amount of the training data from 60 per cent to 80 per cent. We can also observe that the usage of feature discriminator does play a critical role when it comes to improving the performance of the Naive Bayes and Random Forest classifiers. For all the three keywords used, the information gain feature discriminator produces better results than the Gain Ratio attribute.

5. CONCLUSIONS

In this work we present a framework for a structural representation of the opinionated expressions in an online book review forum. This approach needs be applied for online reviews across domains with more versatile data for further observation of how human opinions represented in a particular way can impact. In future using more powerful classifier algorithms as well as usage of more extensive domain test data like twitter would be a viable option to for further improvement of our framework.

6. REFERENCES

[1] Liu, Bing, "Sentiment analysis and subjectivity," Handbook of natural language processing, Volume 2, 2010, pp. 627-666.

[2] Liu, Bing, "Web data mining: exploring hyperlinks, contents, and usage data," Springer Science & Business Media, 2007.

[3] Liu, Bing, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, Volume 5, no. 1, 2012, pp. 1-167.

[4] Wiebe, Janyce M., Rebecca F. Bruce, Thomas P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 246-253.

[5] Arjun Chaudhuri, "Emotion and reason in consumer behavior," Routledge, 2006.

[6] Morariu, I. Daniel, Cretulesku G Radu, Breazu Macarie, Feature selection in document classification.

[7] Zia, Tehseen, Qaiser Abbas, Muhammad Pervez Akhtar, "Evaluation of Feature Selection Approaches for Urdu Text Categorization," International Journal for Intelligent Systems and Applications, 2015.

[8] Meesad, Phayung, Pudsadee Boonrawd, Vatinee Nuipian, "A chi-square-test for word importance differentiation in text classification," In Proceedings of International Conference on Information and Electronics Engineering, 2011, pp. 110-114.

[9] Debole, Franca, Fabrizio Sebastiani. “Supervised term weighting for automated text categorization,” Text mining and its applications, Springer Berlin Heidelberg, 2004, pp. 81-97.

[10] Ikonomakis, M., S. Kotsiantis, V. Tampakas, “Text classification using machine learning techniques,” WSEAS Transactions on Computers, Volume 4, no. 8, 2005, pp. 966-974.

[11] Karegowda, Asha Gowda, A. S. Manjunath, M. A. Jayaram, “Comparative study of attribute selection using gain ratio and correlation based feature selection,” International Journal of Information Technology and Knowledge Management 2, 2010, no. 2, pp. 271-277.

[12] Novovičová, Jana, Antonin Malik, “Information-theoretic feature selection algorithms for text classification,” in Proceeding of IEEE International Joint Conference on Neural Networks, 2005, Vol. 5, pp. 3272-3277.

[13] N. Nicolosi, “Feature selection methods for text classification,” http://www.cs.rit.edu/~nan2563/feature_selection.pdf, 2008.

[14] S. R. Singh, H. A. Murthy, T. A. Gonsalves, “Feature selection for text classification based on gini coefficient of inequality,” Journal of Machine Learning Research - Proceedings Track, vol. 10, 2010, pp. 76-85.

[15] Changxing Shang, Min Li, Shengzhong Feng, Qingshan Jiang, Jianping Fan, “Feature selection via maximizing global information gain for text classification,” Know.-Based Syst. 54 (December 2013), pp. 298-309.

[16] Yiming Yang, Jan O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” In Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Douglas H. Fisher (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412-420.

[17] Jana Novovičová and Antonín Malík and Pavel Pudil, “Feature Selection using Improved Mutual Information for Text Classification”, Lecture Notes in Computer Science, Springer, 2004, pp. 1010-1017.

[18] Monica Rogati and Yiming Yang, “High-performing feature selection for text classification. In Proceedings of the eleventh international conference on Information and knowledge management (CIKM 2002). ACM, New York, NY, USA, pp. 659-661.

[19] Leena H. Patil, Atique Mohammad, “A multistage feature selection model for document classification using information gain and rough set,” International Journal of Advanced Research in Artificial Intelligence(IJARAI), Volume 3 Issue 11, 2014.

[20] Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari, “Feature selection for text categorization on imbalanced data,” SIGKDD Explor. Newsl. 6, 1 (June 2004), pp. 80-89.