

Algorithm to Monitor Suspicious Activity on Social Networking Sites using Data Mining Techniques

Suhas Pandhe

Computer Engineering Department,
Pune Institute Of computer Technology,
Savitribai Phule Pune University, India

Sahil Pawar

Computer Engineering Department,
Pune Institute Of computer Technology,
Savitribai Phule Pune University, India

ABSTRACT

In today's world social networks have become a platform to express opinions or feelings related to current events or any other topics. But sometimes provocative posts related to renowned people, religion, sexuality, countries or any other sensitive topics create havoc in the society. Such posts must be administered and removed before they spread and hurt people's feelings resulting into tension in the society and possible riots. This paper discusses about the techniques to identify such suspicious posts and report them to curb the spread of provoking posts.

Keywords

Data mining, Data Analysis, ID3 Decision Tree

General Terms

Data Mining, Social Media Monitoring

1. INTRODUCTION

Modern applications use huge data bases to store their data which is collected daily. But only some of data is valuable and after processing it we get information important to analyze and take important decisions based on information-rich data. It is similar to mining platinum where kilos of ore are used to extract few grams of precious platinum metal.

Data Mining can be defined as discovering patterns and relationship between data in databases.

This paper will be discussing the following points:-

1. Data Analysis Techniques
2. Monitoring Social Media
3. ID3 Decision Tree
4. Implementation

2. DATA ANALYSIS TECHNIQUE

Data analysis consists of following steps:

2.1 Cleaning and Integration:

The data collected in data warehouse from multiple databases and is filtered to remove unwanted inconsistent data[3].

2.2. Selection and transformation:

In this step relevant data is retrieved from data Warehouse and transformed into appropriate forms using aggregation operations[3].

2.3. Data Mining:

It is a crucial method where different algorithms like C4.5 (decision trees); k-means are applied to extract data patterns[3].

2.4. Pattern Evaluation and Presentation:

In this method various patterns and relationship in between the data set are identify. Then this processed information is represented using bar graphs, other graphical interface[3].

3. RELATED WORK

Very little work has been done on this particular segment of Opinion mining, which is mining Social media for keeping the social world clean and peaceful. Up until now, most research has been done on products and commercial applications. Opinions and comments are generally in form of text.

A social issue is an issue which relates to people's personal lives and interactions [5]. The impact of public opinions about social issues on policy makers operates in a similar manner to those of customer reviews for manufacturers. Governments and organizations who are engaged with social issues can analyze public criticisms or supports for a particular policy and consider public opinions in making decisions. Politicians can receive electorates' opinions concerning important issues and their expectations [5].

4. MONITORING SOCIAL MEDIA

Data mining can be used to monitor social network for suspicious posts or comments. A recent event on the social network triggered the need for such an application [2]. Religious or political sentiments can be hurt with such malicious posts. Social media can be used to broadcast a message to large population quickly. Hence it becomes crucial to monitor the posts on social media. Taking the above situation into account an idea is put forth of an application that a social networking websites like Facebook or twitter would adopt. This application collects the posts and comments from the social networking sites and analyses those post and comments using a customized ID decision tree algorithm.

These posts and comments are analyzed for provoking posts by comparing the words in the posts with the set of sensitive key-words in the algorithm.

Further the set of sensitive keywords are divided into

5 categories:- political, sexuality, religious, historical, racial discrimination. A table to maintain count for each category for all the comments is maintained.

A training data set of 200 comments is given to 3 people to reduce the possibility of individual opinions for analyzing abusive nature of comments. These people rate each of the comment as an intensity factor on the scale of 1 to 5, 1 being least abusive and 5 being most abusive. The average of intensity factor given by all the people is calculated and that becomes the intensity factor for the comment. The same average value is assigned to the category to which the abusive key word belongs Also a threshold value of 3 is set to check if it is abusive or not.

Simultaneously in the comments of data set if the algorithm comes across any of sensitive words related to any of the 5 categories, then it increments the count for that particular category to which the sensitive word belongs. Count for a particular category is the number of keywords matched of that category.

Ex:-It is not true but example of abusive comments would be
Comment 1:-Most of the sex offenders are Muslims.

Comment 2:-Mahatma Gandhi allowed separation of Hindus and Muslims.

Comment 3:- Shivaji Maharaja was the best Maratha king

Comment 4:-Black people are mostly prisoners and slaves taken from Africa a long time ago.

Table 1.Count of keywords matched in a comment

Comment	Religious	Sexual	Racist	Historical	Political
1	1	1	-	-	-
2	1	-	-	-	1
3	-	-	-	1	-
4	-	-	1	-	-

Table 2.Intensity values given by all judges to a comment

Comment	Intensity By Judge1	Intensity By Judge2	Intensity By Judge3	Average Intensity
1	5	5	4	4.67
2	1	2	1	1.33
3	1	1	1	1
4	3	4	2	3

Table 3. Average Intensity value given by all judges for all categories

Comment	Religious	Sexual	Racist	Historic	Political
1	4.67	4.67	-	-	-
2	1.33	-	-	-	1.33
3	-	-	-	1	-
4	-	-	3	-	-
Final Intensity Value	3	4.67	3	1	1.33

In the above example four comments are analyzed for representative purposes.

In each of the four comments a rating is given by 3 experts. The average for particular comment is calculated. In case of comment 1 the average is 4.67 as shown in Table 2. Comment 1 has two keywords –Muslim belonging to Religious category and Sex offenders belonging to Sexual category. So the

average intensity value calculated in Table 2 is now given to Religious and Sexual category as shown in Table 3. Now the Final Intensity for particular category is calculated by taking the average of all values in religious category. In case of Religious category it is 3.

Table 4. Intensity factor for each category after analyzing 200 comments

CATEGORY	INTENSITY
Religious	2.92
Sexual	3.18
Racist	2.47
Political	2.53
Historical	2.31

From the above table a bar graph is plotted.

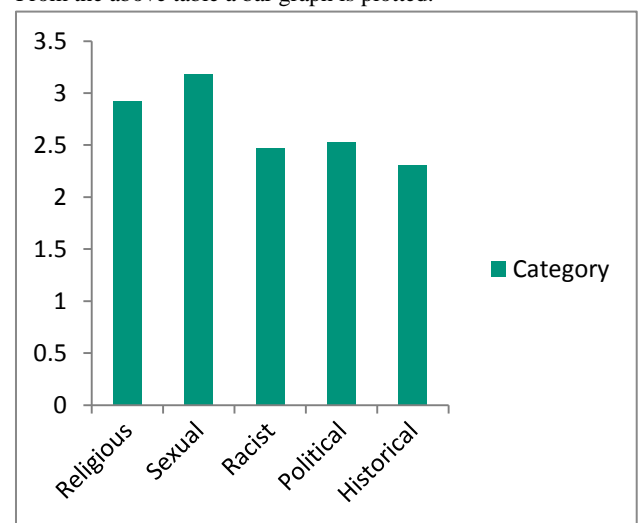


Fig.1-Intensity values for each category after analyzing 200 comments

Figure 1 tells that after analyzing 200 comments the intensity values of sexual and religious are high than rest of the attributes.

These intensity values are used to calculate weight. This weight is calculated using the formula,

Weight of a Comment (W)=(Likes for a Comment)*(Intensity Factor)*(Count of Keywords matched).

For ex:-

Comment:-Most of the sex offenders are Muslims.
Likes=60

In this Comment 1 keyword each belonging to Religious and Sexual Category have been found.

Weight for Sexual category:-

Count=1

Intensity=3.18

$W1=3.18*1*60=190.8$

Weight for Religious category:-

Count=1

Intensity=2.92

$$W2=2.92*1*60=175.2$$

Total Weight of Comment=W1+W2=366

So the algorithm will create the Table 5 for all comment:-

Table 5-Weights for all categories for each comment

Comment	TW	Re	S	Ra	H	P
1	366	175.2	190.8	-	-	-

TW=Total Weight

Re=Religious

S=Sexual

Ra=Racial

H=historical

P=Political

5. ID3 DECISION TREE

Iterative Dichotomiser 3 or ID3 is an algorithm which is used to generate decision tree.ID3 decision tree has been used to classify the comments into abusive and non-abusive.

First a table is created which contains comment number versus all 5 categories for a training data set. The basic step in constructing ID3 decision tree is attribute selection. The two mathematical terms used to process attribute selection are Entropy^[3] and Information Gain^[3]. These 2 terms will be used by the ID3 algorithm to determine which attribute will be selected to become a node of decision tree. All the records in the table are referred as the Collection(c).

Calculation of Entropy:

Entropy is defined as the average amount of the information contained in the event, sample or record.

$$\text{Entropy}(S) = \sum_{n=1}^n p(I) \log_2 p(I)$$

p(I) refers to the proportion of S belonging to class I. \sum is over c i.e. summation of all the classifier items.

Information Gain

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum ((|S_v|/|S|) \times \text{Entropy}(S_v))$$

Where,

S is the total collection of the records.

A is the attribute for which gain will be calculated.

v is all the possible of the attribute A

S_v is the number of elements for each v for instance \sum is the summation of ((|S_v|/|S|) x Entropy(S_v)) for all the items from the set of v

6. IMPLEMENTATION

This concept is implemented using ID3 decision tree algorithm. An example of experimental training dataset is shown in Table 5. There is an entry for each post in this table. The second column is total weight.

The total weight is a way of representation of circulation of sensitive posts over social networking as it considers all the factors like likes and shares of comments, intensity of keywords matched and count of keywords.

Table 6-Total weight is represented as high, low medium values while all other categories are represented as high and low

COMMENT	TOTAL RATE/COUNT	RELIGIOUS	SEXUAL	RACIST	POLITICAL	HISTORICAL	SUSPICIOUS
1	HIGH	HIGH	LOW	LOW	LOW	LOW	YES
2	MEDIUM	HIGH	HIGH	LOW	LOW	LOW	YES
3	LOW	LOW	HIGH	HIGH	HIGH	LOW	NO
4	MEDIUM	LOW	LOW	HIGH	HIGH	HIGH	NO
5	HIGH	LOW	LOW	HIGH	LOW	LOW	YES
6	LOW	HIGH	LOW	LOW	HIGH	HIGH	NO
7	HIGH	HIGH	LOW	HIGH	LOW	LOW	YES
8	MEDIUM	HIGH	LOW	LOW	LOW	LOW	YES
9	LOW	HIGH	HIGH	LOW	LOW	LOW	YES
10	HIGH	LOW	HIGH	LOW	LOW	LOW	YES
11	MEDIUM	LOW	HIGH	LOW	LOW	HIGH	NO
12	LOW	HIGH	HIGH	HIGH	LOW	LOW	YES
13	MEDIUM	HIGH	LOW	LOW	LOW	LOW	YES
14	LOW	HIGH	LOW	HIGH	LOW	HIGH	NO
15	HIGH	LOW	LOW	LOW	LOW	LOW	YES

For total weight category 2 thresholds t_1 and t_2 have been defined such that it will represent the total weight into ternary values high, low and medium.
Let weight be w then:-

If $w < t_1$ then value=low
 $t_1 < w < t_2$ then value=medium
 $w > t_2$ then value=high.

For religious, sexual, racist, historical categories one threshold i.e. t_3 is defined that will represent all these categories into binary values low and high.

Let w_1 be individual weight of religious category then:-
If $w_1 < t_3$ then value=low
 $w_1 > t_3$ then value=high.

Similarly it can be applied to the remaining individual categories racial, sexual, political, and historical
After analyzing 200 comments, a table is generated in which all the values of weights are assigned a value that is high or low or medium. For example a sample table of 15 comments has been shown in Table 6 :-

Id3 will use the Table 6 to calculate two terms namely Entropy and Information gain.
 $Entropy(S) = \sum -p(I)\log_2 p(I)$

$p(I)$ refers to the proportion of S belonging to class I i.e. in the above table there are two kinds of classes:- {No, Yes} with {5, 10} (in here 5 is the total number of No and 9 is the total no of Yes), the collection size is $S=15$. So the $p(I)$ over C for the Entire collection is No (5/15) and Yes (10/15).

$$S = -5/15 * \log_2(5/15) - 10/15 * \log_2(10/15)$$

$$S = 0.9183$$

$Gain(S, A) = Entropy(S) - \sum ((|S_v|/|S|) \times Entropy(S_v))$
Then Gain is calculated as shown below,

$$\text{Gain}(S, \text{Religious}) = Entropy(S) - ((|S_{low}|/|S|) \times Entropy(S_{low})) - ((|S_{high}|/|S|) \times Entropy(S_{high}))$$

The Entropy (S_{low}) and Entropy (S_{high}) are calculated as follows:-

For Religious category the number of high, low is (9, 6)

For low the number of (yes, no) is (3, 3)

For high the number of (yes, no) is (6, 3)

$$Entropy(S_{low}) = -(3/6) * \log_2(3/6) - 3/6 \log_2(3/6)$$

Therefore, $S_{low}=1$

$$Entropy(S_{high}) = -(7/10) * \log_2(7/10) - 2/10 \log_2(2/10)$$

Therefore, $S_{high}=0.824$

Now there is S_{high} and S_{low} for religious category, Gain of religious column is calculated as:-

$$\text{Gain}(S, \text{Religious}) = Entropy(S) - ((|S_{low}|/|S|) \times Entropy(S_{low})) - ((|S_{high}|/|S|) \times Entropy(S_{high}))$$

$$\text{Gain}(S, \text{Religious}) = 0.9183 - (6/15) * 1 - (9/15) * 0.824$$

$$\text{Gain}(S, \text{Religious}) = 0.02366$$

Similarly for all other attributes like sexual, political, racist and historical the information gain will be calculated as above.

7. RESULTS

For the above application Id3 decision tree is constructed. After executing the Id3 algorithm the following results were obtained. The input to this application is a text file containing a table for 200 comments and the number of attributes.

```

<terminated> Id3_1 [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe
ENTER THE NAME OF FILE FOR TRAINING:- comment.txt
ENTER THE NUMBER OF ATTRIBUTES:- 6
WHEN TOTAL IS "HIGH"
        SUSPICIOUS = "YES".
ELSE WHEN TOTAL IS "MEDIUM"
        WHEN RELIGIOUS IS "HIGH"
                SUSPICIOUS = "YES".
        ELSE WHEN RELIGIOUS IS "LOW"
                SUSPICIOUS = "NO".
ELSE WHEN TOTAL IS "LOW"
        WHEN SEXUAL IS "LOW"
                SUSPICIOUS = "NO".
        ELSE WHEN SEXUAL IS "HIGH"
                WHEN RELIGIOUS IS "HIGH"
                        SUSPICIOUS = "YES".
                ELSE WHEN RELIGIOUS IS "LOW"
                        SUSPICIOUS = "NO".
    
```

Fig.2-Results produced by Id3 Decision tree when Table-6 is given as input

When the table is given to the Id3 algorithm it produces the following decision tree as shown in Figure 3. After calculating the information gain for all the attributes, the information gain for the total weight comes out to be high. Hence the total weight is taken as the root node. For every node the information gain is calculated recursively till all the leaf nodes are created.

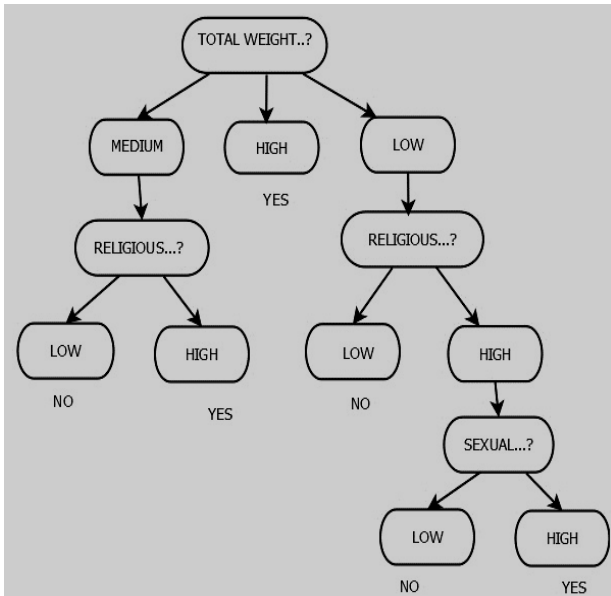


Fig.3-Representation of the Fig. 2 in the form of Decision tree

The above decision tree given in Figure 3 will be created and stored in the databases of all the servers of social networking sites like Facebook and Twitter. When a comment is posted on any social networking site, the words in the comments will be compared with abusive keywords in our algorithm. Once a abusive keyword is matched then all the weights for the attributes will be calculated and a table will be created. This table will be matched will the already created decision tree and if it satisfies the condition for abusive in the decision tree it will be filtered out and comment will be treated as abusive.

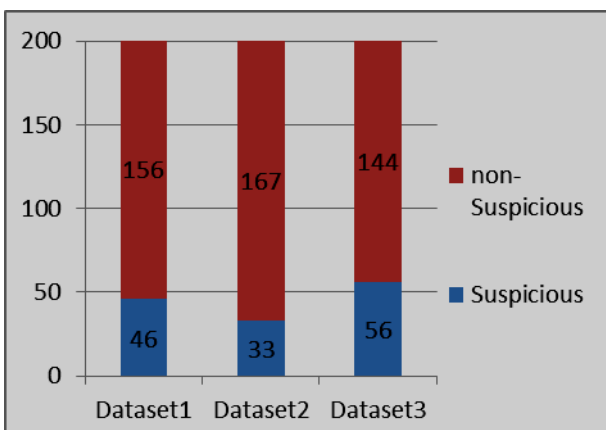


Fig.4-Number of Suspicious and Non-Suspicious comments from three training dataset of 200 comments

The Figure 4-shows us the number of suspicious and non-suspicious comments for 3 data sets each containing 200 comments. On an average for every 200 comments 45 suspicious comments were found. Most of these suspicious comments had high content of religious keywords.

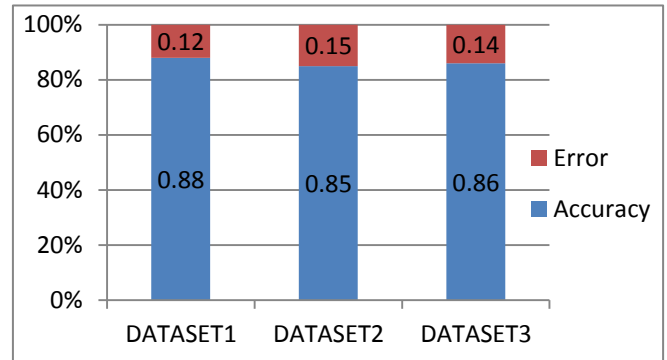


Fig.5- Accuracy for three training dataset of 200 comments

Using Id3 decision tree highly accurate results are obtained as can be seen in Figure 5. The contributions to this high accuracy are the two mathematical terms Entropy and Information Gain. On an average an accuracy of 86% is achieved which is relatively high compared to neural networks and other classification algorithms. The 14% error is due to some of contradictory comments which contains abusive words but are not actually abusive comments.

4. CONCLUSION

We have presented an application of data mining which will help to curb the effect of posting abusive articles or comments on social network that is likely to harm the sentiments of society knowingly or unknowingly. This algorithm is based on ID3 Decision tree. The preprocessing is done by customized algorithm which will give a table having comment number and its value for all the categories like religious, sexual political, historical and racist. This table is then given to ID3 decision tree algorithm which will give a decision tree which will monitor the derogatory comments on social networking sites. The results of this algorithm are about 86% accurate.

On adoption of the algorithm by social networking sites, it will monitor social behavior and derogatory posts can be deleted. In future we plan to integrate the algorithm with social networking sites like Facebook and Twitter and do the necessary modifications in algorithm to suite their databases and to able to handle the processing of large amount of comments every day.

5. REFERENCES

- [1] S. Somasundaran and J. Wiebe. Recognizing stances in ideological online debates. In Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 116–124. ACM, 2010
- [2] Mostafa Karamibekr and Ali A. Ghorbani, Sentiment Analysis of Social Issues. In 2012 International Conference on Social Informatics, Pages 215 – 221. IEEE, 2012
- [3] Data Mining: Concepts and Techniques by Han and Kamber [Online].
- [4] <http://www.codeproject.com/Articles/259241/ID-Decision-Tree-Algorithm-Part>
- [5] Managem MacDougall, R. (2005), Identity, electronic ethos, and blogs: a technological analysis of symbolic exchange on the new news medium, American Behavioral Scientist, Vol. 49, No. 4, 575–599.

- [6] T.Wilson,., J. Wiebe,., R. Hwa,., 2004. Just how mad are you? Finding strong and weak opinion clauses. In: the Association for the Advancement of Artificial Intelligence, pp. 761--769.
- [7] W. Jin, H. Hay Ho, and R. Srihari, 2009. Opinion Miner: A Novel Machine Learning System for Web Opinion Mining and Extraction. Proceeding of International conference on Knowledge Discovery and Data Mining Paris, France.
- [8] Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. Mining newsgroups using networks arising from social behavior. In Proceedings of WWW, pages 529–535, 2003.
- [9] Gharehchopogh, F.S, Khalifelu, Z.A, 2011, Application Data Mining Methods for Detection Useful Knowledge in Health Center: A Case Study Using Decision Tree, International Conference on Computer Applications and Network Security (ICCANS 2011), 1-5.
- [10] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005
- [11] Usama Fayyad;Gregory Piatetsky-Shapiro and Padhraic Smyth,From Data Mining to Knowledge Discovery in Databases.In American Association for Artificial Intelligence,2006
- [12] Aggarwal, C.: An introduction to social network data analytics. Springer US, 2011
- [13] Liu, B.: Sentiment analysis and opinion Mining. AAAI-2011, San Francisco, USA, 2011
- [14] Ku, L.-W., Liang, Y.-T., Chen, H.-H.: Opinion extraction, summarization and tracking in news and blog corpora. In Proc. of the AAAI-CAAW'06, 2006