SENTIMENT ANALYSIS TECHNIQUES TO ANALYZE HSE SITUATIONAL AWARENESS AT OIL AND GAS PLATFORMS USING MACHINE LEARNING

Dafuallah Esameldien Dafaallah
Computer Information Science Department, Universiti Teknologi PETRONAS
Bandar Seri Iskandar, Perak, 32610, Malaysia
dafuallah_18003577@utp.edu.my

Ahmad Sobri Hashim
Computer Information Science Department, Universiti Teknologi PETRONAS
Bandar Seri Iskandar, Perak, 32610, Malaysia
sobri.hashim@utp.edu.my

Abstract - Health Safety & Environment (HSE) situational awareness is a very important aspect of any risky workplace. Negligence in complying with HSE policies and practices might lead to unwanted incidents, critical injuries, death, spread of diseases and environmental pollution. In most corporations, information on HSE related incidents is disseminated through formal channels such as reports. Employees on the other hand frequently use social media to share, complain and discuss HSE-related issues. The issues are discussed through an informal platform, it is difficult to analyze opinions for further action. Therefore, this study will investigate existing sentiment analysis models and formulate a suitable sentiment analysis model using machine learning technique. Through literature review, Naïve Bayes model was found to be the most efficient text classification in sentiment analysis. This technique still needs further enhancement as the accuracy is not within requirement. Upon enhancing the Naïve Bayes model, a better outcome can be attained.

***Keywords*:** HSE; Sentiment Analysis; Naïve Bayes; Machine Learning

1. Introduction

Situational awareness has always been an important aspect in controlling, minimizing and avoiding health, safety and environment (HSE) related incidents in the workplace. It refers to the act of employees being aware of their surroundings in terms of where they are, where they are supposed to be, and if anyone or anything around them is a threat; and subsequently evaluate the risk for accidents to happen (Sneddon et al., 2004), (Endsley, 2015), (Mrema et al., 2015). Sentiment analysis (SA) on the other hand is a popular automated computational process to analyze and understand the opinions discussed on social media platforms (Farhadloo & Rolland, 2016). This process uses information technology (IT) capabilities to extract information on certain issues discussed through social media platforms (Kietzmann, 2016). The analysis of information extracted will then produce a series of recommended actions to be considered by the organization. In the oil and gas industry, the current process of decision making for HSE-related issues is based on formal reports prepared through scheduled inspections, or informal reports made by employees via email or common communication channels. This process of decision making consumes a longer time as it needs to follow the standard investigation procedures. Furthermore, employees usually share or comment on everything they perceive and experience regarding HSE-related issues through social media. Using SA, information can be gathered from social media and analyzed to assist in the decision making for HSE-related incidents. There have been many existing models developed using various techniques to analyze sentiments on social media such as to detect terrorism via Naïve Bayes technique (Azizan & Aziz, 2017), observing citizen's perception towards government's performance (Azizan & Aziz, 2017), assessing rating of certain product using Support Vector Machine (Sekharan, 2017), reviewing news articles using Lexicon-based technique (Vohra & Teraiya, 2013), understanding lexicon and learning-based approach (Mudinas et al., 2012, August), (Zhang et al., 2011). However, SA for the HSE domain has yet to be studied upon. Existing SA models also may not be suitable for this study due to the difference in measured parameters. Therefore, this study will investigate existing SA techniques and algorithms used to analyze expressed sentiments in social media. Finally, the accuracy of the formulated model in analyzing sentiments of HSE-related incidents on social media

will be validated. It is expected that the current research will contribute towards the improvement of machine learning and the formulated SA model will bring about a big improvement in the SA implementations. In accordance with the underlying concept described in this section, the rest of the paper would discuss, in order, the literature review on existing SA methods, results of the proposed algorithm performance followed by concluding remarks and directions for future works.

2. Literature Review

2.1. *Health, Safety and Environment*

HSE has always been a crucial aspect in any workplace. It is concerned with preparing a workplace that can minimize or avoid the occurrence of incidents which affects an employee's wellbeing (Friend & Khon, 2007). The standard definition of HSE involves a framework that includes several components such as family, health, sustainability, society, environment and economy (Molamohamadi & Ismail, 2014). HSE policies, frameworks and practices can vary from one company to another based on their businesses. Most companies worldwide have their own HSE policy which is adopted or adapted from international policies guided by the international HSE agency, Occupational Safety and Health Administration (OSHA) (Molamohamadi & Ismail, 2014). In Malaysia, the agency who is responsible and enforces the HSE related acts and guidelines is Department of Occupational Safety and Health (DOSH), established under the Ministry of Human Resource. Employees in the oil and gas sector especially in upstream are regularly exposed to hazards and they are bound to many HSE regulations and guidelines. The incidents are not only harmful towards a single employee, as it has a risk to affect other employees within the workplace environment. For example, the Deepwater Horizon incident which occurred on April 20, 2010, caused the loss of eleven lives and another seventeen employees were seriously injured (Dragos, 2011). Not only that, this disaster has also caused an oil spill which released dangerous chemicals into the ocean and atmosphere (Austin et al., 2014). Other than that, employees in the oil and gas industry are also exposed to other risks of incidents involving transportation, machinery and heavy equipment, fire and explosion, chemical and environment, and falling from high places. Oil and gas employees undergo thorough safety training to avoid and minimize the risks of such incidents from happening. Besides, there have been existing preventive actions practiced such as routine inspection, regular HSE reporting, etc. However, the safety information is only limited to HSE personnel and other employees who reported the particular incident.

2.2. *Sentiment Analysis Techniques*

SA techniques have been given high attention since the influx in the era of the Internet. Pang et. al. (Pang et al., 2002, July) also considered the problem of classifying by overall sentiment instead of topic to determine whether a review is positive or negative. They also proposed three algorithms on measuring a sentiment positive or negative review.

(i) Naïve Bayes (Pang et al., 2002, July)

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

(ii) Maximum Entropy (Pang et al., 2002, July)

$$Pme(c|d) := \frac{1}{Z(d)} \exp = (\sum_{i} \lambda_{i,c} F_{i,c}(d,c))$$

(iii) Support Vector Machine (Pang et al., 2002, July)

$$w := \sum_{j} a_j c_j d_j, \quad d_j \geq 0$$

In another study, Pang and Lee (Pang & Lee, 2008) tested the algorithms in sentimental education, where SA will be used for subjective summarization based on minimum cuts in the graph. Based on these studies, it can be assumed that; SA could be used for more than just a movie review and for finding minimum cuts in the graph. Furthermore, Vohra and Teraiya (Vohra & Teraiya, 2013) discussed SA's contribution towards the current growth of social media. They used SA to analyze bulk data generated by users on the social web. Examples of the data which they have used are reviews of comments and opinion to analyze positivity, negativity and neutral reactions of users by building an intelligent system.

2.3. *Machine Learning Technique*

Machine learning approach applicable to SA mostly belongs to supervised classification. In machine learning-based techniques, two sets of documents are needed: training set and test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents whereas a test set is used to check how well the classifier performs. Several machine learning techniques have been adopted to classify the reviews.

(i) Naïve Bayes

Several methods exist to perform text classification. One of the effective approaches is to assign to a given document, d the class c∗ = arg maxc P (c | d). The Naïve Bayes' formula is as follows: (Pang et al., 2002, July)

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

The simplicity of this formula's conditional independence assumption clearly does not hold in real-world situations. Despite this, Naive Bayes is optimal for certain problem classes with highly dependent features. On the other hand, more sophisticated algorithms might yield better results (Sekharan, 2017), (Annett & Kondrak, 2008, May), (Medhat et al., 2014).

(ii) Maximum Entropy

Maximum entropy is an alternative technique which has proven effective in several natural language processing applications. At times, it outperforms Naive Bayes at standard text classification. Naïve Bayes performs better when it comes to standard text classification. The Maximum Entropy formula is as follows: (Pang et al., 2002, July)

$$Pme(c|d) := \frac{1}{Z(d)} \exp = (\sum_{i} \lambda_{i,c} F_{i,c}(d, c))$$

For instance, a particular feature/class function might fire only if the bigram "still hate" appears and the document's sentiment is hypothesized to be negative. Importantly, unlike Naive Bayes, maximum entropy makes no assumptions about the relationships between features, and potentially perform better when conditional independence assumptions are not met. The $\lambda_{i,c}$'s are feature-weight parameters; inspection of the definition of PME shows that a large $\lambda_{i,c}$ means that f_i is considered a strong indicator for class c. The underlying philosophy is that we should choose the model that makes the fewest assumptions about the data while still remaining consistent with it, which makes intuitive sense (Chen et al., 2016).

(iii) Support Vector Machine

Moreover, Naïve Bayes formula can be compared with support vector machines formula. Support vector machines (SVMs) are known to be highly effective at traditional text categorization, generally outperforming Naive Bayes. SVM Formula can be derived according to the following figure (Vohra & Teraiya, 2013), (Pang et al., 2002, July).

$$w := \sum_{j} a_j c_j d_j, \quad d_j \geq 0$$

Pang (Pang et al., 2002, July) has made a comparison at document level in sentiment classification by comparing the performance of three machine learning classifiers which are Naïve Bayes, Maximum Entropy and Support Vector Machines on different features like considering combining unigrams, bigrams, combination of both or only unigrams. The result has shown that feature presence is more important than feature frequency. When the feature to be extracted is a small set, Naïve Bays performs better than SVM. However, Pang (Pang et al., 2002, July) also proved that if feature space is increased, SVM's performed better and Maximum Entropy may perform better than Naïve Bayes but may suffer from overfitting. In terms of different languages such as English and Arabic, Abbasi (Abbasi et al., 2008) proposed SA on the forte of hate/extremist web forum by utilizing stylistic and syntactic features. Entropy Weighted Genetic Algorithm is a hybrid genetic algorithm that uses the information gain heuristic method to improve feature selection. They used Support Vector Machine (SVM) with 10- fold cross-validation and bootstrapping to classify sentiments in all experiments. When using both syntactic and stylistic features, they achieved 95.55% accuracy in 10 crosses validation.

Table 1. Machine Learning Technique Comparison

Techniques/ parameters	**Support Vector Machine**	**Maximum Entropy**	**Naïve Bayes**
Accuracy	High	Low	Medium
Training Time	High	High	Low
Advantages	Can be better cope in many noisy features, Hard to interpret the data	Easy to use with fewer parameters.	High scalability of data and fast classification rate
Disadvantages	Long Training Time Needed	Long training time needed	Low accuracy if using less training data

3. Methodology

3.1. *Research Activities*

(i) Phase 1

Interview method has been conducted to identify business requirements with the business experts and target users which are managers of HSE division from one of the Oil and gas companies in Malaysia. This activity has proceeded with the investigation of the existing SA techniques and algorithms used to analyze sentiments expressed in social media with regards to HSE. The purpose of this investigation is to identify the parameters and variables that have been used to analyze the sentiments, and the accuracy of the applied techniques and algorithms in analyzing the sentiments. Existing techniques and algorithms were then studied thoroughly to understand how each technique under machine learning SA works. From the analysis and comparison, a decision can be made to select the appropriate method for this research and requirements collected.

(ii) Phase 2

The annual historical dataset has been supplied by an undisclosed oil and gas company in the format of CSV file. Discussion, as well as meetings, have been conducted with field experts to better understand the supplied datasets and familiarize with the overall process of various data form collection, and the expected behavior of the oil and gas companies. From here, data cleansing could also be done so that missing values, as well as noise, could be removed from the dataset. The cleansed dataset was separated into 2 sets of data, mainly training dataset and testing dataset. Training dataset has taken up 80% of the whole dataset where this is used to train the algorithm while the remaining 20% is used as the test dataset during validation and comparison of training result. The dataset is processed via word embedding to create vectors that are able to have higher accuracy when it comes to determining the polarity of the sentence. Word embedding is a type of word presentation that allows words with similar meaning to have the same presentation. Word2Vec is used here as it is a statistical method for efficiently learning a standalone word embedding from a text corpus. This SA technique was used for this research as it is best suited when it comes to accuracy and predictability.

4. Result and Discussion

In this study, since the data for HSE has not been received from the relevant department, it has been replaced with customer reviews of any random product. The data was scraped from Amazon for a review on a product. These reviews were commented on by multiple customers. The reviews are used in this case as it is similar to the data taken from Twitter on HSE. People's opinions, thoughts on an incident or product are similar to unfiltered comments from netizens on social media as it contains keywords such as "bad", "dangerous" and other opinions given by customers.

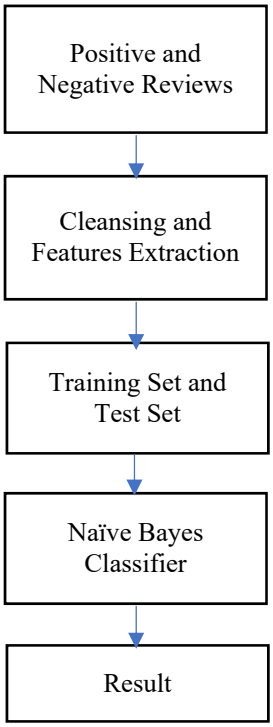

Figure 1. Sentiment Analysis Technique using Naïve Bayes

In Figure 1, the training datasets includes positive and negative reviews. Before it can be classified, cleansing and features extraction is done to ensure that no anomalies disrupt the training process. The training process is completed, and test sets are used to evaluate the accuracy of the Naïve Bayes technique by dividing the output number of positive/negative test sets over the actual positive/negative test sets. The data of customers reviews was used, and it was sourced online. However, the method of collecting such reviews is by using a scrapper that scrapes data from a website and stores it in CSV format. Once exported, the data then is simplified and labelled according to 2 categories. _label 1 refers to customers having 1-3-star reviews which is negative, and _label 2 refers to 4-5-star reviews which are positive.

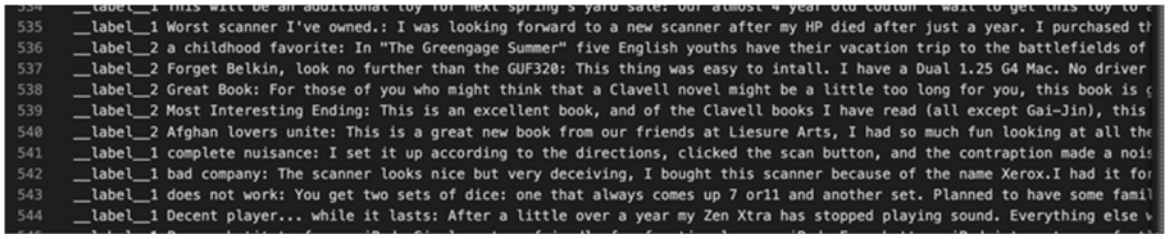

Figure 2. Data Sample

The data were split into training and test sets. 20% test sets and 80% training sets. Before feeding into the model, the first thing that was done is to ensure the data set is in the correct format since the format file it captures is in .bz2 and was converted to strings that are passed. Figure 3 shows a sample of raw data.

```
'Stuning even for the non-gamer: this sound track was beautiful! It paints the senary in your mind so well I would
recomend it even to people who hate vid. gAme music! I have played the game chrono cross but out of all the game I
have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate
guit ars and soulful orchestras. It would impress anyone who cares to listen! ^_^''
```

Figure 3. Raw Data

In Figure 3, cleansing and feature extraction is done to ensure the data is cleansed and does not have any unwanted symbols or formats. Firstly, the sentence letters are reduced to lower case, called case folding.

```
'stuning even for the non-gamer: this sound track was beautiful! it paints the senary in your mind so well i would
recomend it even to people who hate vid. game music! i have played the game chrono cross but out of all the game i
have ever played it has the best music! it backs away from crude keyboarding and takes a fresher step with grate
guit ars and soulful orchestras. it would impress anyone who cares to listen! ^_^''
```

Figure 4. Case Folded Sentence

Secondly, special characters or non-alphabetical characters were removed from the sentence to ensure unwanted data.

```
 stuning even for the non gamer  this sound track was beautiful  it paints the senary in your mind so well i would
recomend it even to people who hate vid  game music  i have played the game chrono cross but out of all the game i
have ever played it has the best music  it backs away from crude keyboarding and takes a fresher step with grate
guit ars and soulful orchestras  it would impress anyone who cares to listen
```

Figure 5. Non-Alphabetical Characters are removed

Thirdly, stop words were removed from the sentence as they don't have any meaning to the sentence. Example of stops words are "the", "a", "how" or "what".

```
['stuning', 'even', 'gamer', 'sound', 'track', 'beautiful', 'paints', 'senery', 'mind', 'well', 'would', 'recomend',
'even', 'people', 'hate', 'game', 'music', 'played', 'game', 'chrono', 'cross', 'games', 'ever', 'played', 'best', 'm
usic', 'backs', 'away', 'crude', 'keyboarding', 'takes', 'fresher', 'step', 'grate', 'guitars', 'soulful', 'orchestra
s', 'would', 'impress', 'anyone', 'cares', 'listen']
```

Figure 6. Stop words removed

In order to create a classifier, a feature extractor was used to extract and sort the features according to its relevance. In this case, the input is the test reviews. The word features list was used along with the input to create the dictionary. Test set and data set were then categorized with the labels as 1 and 0 where label 1 is positive sets and label 2 are negative sets. Once the model is trained, the test set is used to determine the ratio of positive and negative for both test set and negative set. The results for positive is 53/40 and negative is 47/42. Converting the result in percentage, the accuracy level is 72%. The Bayes' theorem was applied to predict a class for any given text from the customer review. Below is the formula to be adopted:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

From the formula, P(c) is the extracted data text and can be categorized as positive, negative or neutral while P(d) is the customer review. P(c|d) is the result of this technique. This research result is still in progress as the accuracy percentage has not been achieved.

5. Conclusion

Based on the Literature Review conducted, Naïve Bayes has been found as the most efficient text classification in Sentiment Analysis. However, regarding the experiment conducted, the accuracy of the algorithm still needs improvements to achieve the target. The next step to this research is to modify its variable and apply classifiers to enhance the model's classification.

6. Acknowledgements

This work is supported by Universiti Teknologi PETRONAS Foundation (YUTP) with Ref. No. 015LC0-290.

7. References

[1] Abbasi, A., Chen, H. M., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS), 26(3), 1-34.
[2] Annett, M., & Kondrak, G. (2008, May). A comparison of sentiment analysis techniques: Polarizing movie blogs. In Conference of the Canadian Society for Computational Studies of Intelligence, 25-35.
[3] Austin, D., Dosemagen, S., Marks, B., McGuire, T., Prakash, P., & Rogers, B. (2014). Offshore Oil and Deepwater Horizon. Social Effects on Gulf Coast Communities Key Economic Sectors, NGOs, and Ethnic Groups 2.
[4] Azizan, S. A., & Aziz, I. A. (2017). Terrorism Detection Based on Sentiment Analysis using Machine Learning. Journal of Engineering and Applied Sciences, 12(3), 691-698.
[5] Chen, H. M., Franks, P. C., & Evans, L. (2016). Exploring Government Uses of Social Media through Twitter Sentiment Analysis. Journal of Digital Information Management, 14(5), 290-301.
[6] Dragos, I. N. (2011). Deepwater Horizon disaster and influence on offshore industry regulations. Journal of Engineering Studies and Research, 17(17), 94-101.
[7] Endsley, M. R. (2015). Situation Awareness Misconceptions and Misunderstandings. Journal of Cognitive Engineering and Decision Making, 9(1).
[8] Farhadloo, M., & Rolland, E. (2016). Fundamentals of Sentiment Analysis and Its Applications Social Media and News Sentiment Analysis for Advanced Investment Strategies. 639(9), 1-24.
[9] Friend, M. A., & Khon, J. P. (2007). Fundamentals of Occupational Safety and Health. The Scarecrow Press, 4.
[10] Kietzmann, J. (2016). Crowdsourcing: A revised definition and introduction to new research. Business Horizons, 6(2), 1-3.
[11] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.
[12] Molamohamadi, Z., & Ismail, N. (2014). The Relationship between Occupational Safety, Health, and Environment, and Sustainable Development: A Review and Critique. International Journal of Innovation, Management and Technology, 5(3), 198-202.
[13] Mrema, E. J., Ngowi, A. V., & Mamuya, S. H. D. (2015). Status of Occupational Health and Safety and Related Challenges in Expanding Economy of Tanzania. Annals of Global Health, 81(4), 538-547.
[14] Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, 1-8.
[15] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135.
[16] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques [Association for Computational Linguistics.]. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, 79-86.
[17] Sekharan, S. C. (2017). Sentiment Analysis Based Product Rating Using Textual Reviews. International Conference on Electronics, Communication and Aerospace Technology ICECA 2017, Coimbatore.
[18] Sneddon, A., Mearns, K., Flin, R., & Bryden, R. (2004). Safety and Situation Awareness in Offshore Crews. The Seventh SPE International Conference on Health, Safety, and Environment in Oil and Gas Exploration, Calgary, Alberta, Canada. .
[19] Vohra, S. M., & Teraiya, J. B. (2013). A comparative study of sentiment analysis techniques. Journal JIKRCE, 2(2), 313-317.
[20] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. HP Laboratories, Technical Report HPL, 2011(89).