How to cite this paper:

Zuraini Zainol, Puteri N.E. Nohuddin, Tengku A.T. Mohd, & Omar Zakaria. (2017). Text analytics of unstructured textual data: A study on military peacekeeping document using R text mining package in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 1-7). Sintok: School of Computing.

TEXT ANALYTICS OF UNSTRUCTURED TEXTUAL DATA: A STUDY ON MILITARY PEACEKEEPING DOCUMENT USING R TEXT MINING PACKAGE

Zuraini Zainol[1], Puteri N.E. Nohuddin[2], Tengku A.T. Mohd[3] and Omar Zakaria[4]

[1]Cyber Security Centre, Universiti Pertahanan Nasional Malaysia, Malaysia, zuraini@upnm.edu.my
[2]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia, puteri.ivi@ukm.edu.my
[3]University of Iowa College of Engineering, United State of America, alyamohd2102@gmail.com
[4]Computer Science Department, Universiti Pertahanan Nasional Malaysia, Malaysia, omar@upnm.edu.my

ABSTRACT. This paper describes a technique of text analytics on peacekeeping documents to discover significant text patterns exist in the documents. These documents are considered as unstructured textual data. The paper proposes a framework that consists of 3 stages (i) data collection (ii) document preprocessing and (iii) text analytics and visualization. The technique is developed using R text mining package for text analytics experiments.

Keywords: *Text analytics, unstructured, textual, R mining, peacekeeping*

1. INTRODUCTION

Advanced computer technologies give a convenient platform for us to digitize documents, produce web documents and electronic documents. Therefore, these documents can be managed efficiently and effectively for keeping, copying and reproducing documents and knowledge. Many document management techniques are developed to improvise methods for handling and analyzing content of documents. Text mining is used for discovering useful patterns and knowledge in unstructured text documents.

The aim of this paper is to discover a technique of analyzing the unstructured text data in discovering the useful knowledge from military peacekeeping documents. The authors proposed a framework for Text Analytics of Unstructured Data using R text mining package.

The rest of this paper is organized as follows. Section 2 explores Knowledge discovery, text analytics and military explicit knowledge and some related work on related topics. Section 3 discusses the framework for the proposed technique. Section 4 discusses the experiment and results. Finally, we conclude this paper with future work in Section 5.

2. BACKGROUND AND RELATED WORK

In this section, some background information on unstructured data, text analytics and military explicit knowledge in peacekeeping operations is discussed.

2.1 Unstructured Textual Data

Massive data are generated daily through sales transactions, web hosting, social media either on internet or closed network systems. Substantial growth in the volume is due to the accumulation of unstructured text data. Unstructured text data is considered as textual information that either does not have a pre-defined data format or data schema (Doan *et al*, 2009). Corporations accumulate massive amounts of documents, emails, social media, and other text-based information to gain competitive advantage in their businesses. Text analytics grasps the crucial process for revealing the business value within companies' data assets. The businesses invest appropriate platforms to sufficiently utilize their databases and applied of the up-to-date text analytics and Natural Language Processing (NLP) algorithms.

2.2 Knowledge Discovery using Text Analytics

According to Feldman and Dagan (1995), Knowledge Discovery in Text (KDT) is a process of identifying valid, novel, potentially useful and ultimately understandable patterns in unstructured text data. Text Mining (TM) or Text Analytics (TA) is frequently used for analyzing the unstructured text documents in search of useful information and knowledge hidden from text resources. TA is used in healthcare for investigating patient health outcomes and providing clinical decision making for health practitioners (Massey, 2015). In business world, useful information discovered from TA can disclose interesting patterns or semantic relationship and trends in large volumes of unstructured data. TA is an extension of data mining, which involves multiple disciplines areas such as information retrieval (IR), Statistics, Web Mining, Computational Linguistics and NLP. TA can also be described as intelligent text analysis, text data mining and knowledge discovery in text.

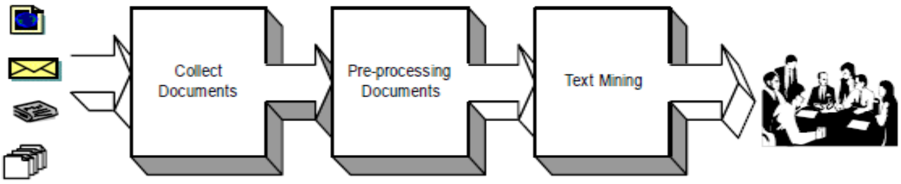


Figure 1. Overview of three main steps in TM process (Karanikas *et al*., 2009)

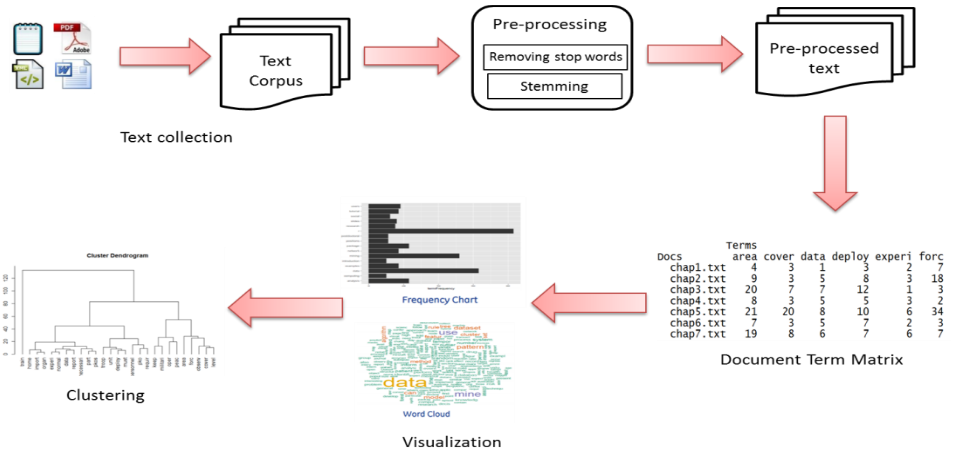
Figure 1 depicts three (3) main stages (i) document collections such as emails, online news, reports, manuals and many more. These documents are usually can be found in the form of semi-structured or unstructured content, (ii) document pre-processing in which document collections are processed and transformed into suitable format for text analysis and (iii) text mining operations where interesting patterns and knowledge are extracted using text mining techniques. These patterns and knowledge are presented to users, which can be used for assisting decision makers in any application domains.

2.3 Military Explicit Knowledge – Peacekeeping Manual

Knowledge is widely associated to processed information or skills acquired during a learning process. Knowledge that has been documented is called as explicit and knowledge based on undocumented lesson learned and experiences are known as tacit knowledge. Similarly, in Military domain, it also can be categorized into (i) explicit knowledge and (ii) tacit knowledge. In the military context, the explicit military knowledge is discussed as knowledge that can easily articulated, codified, accessed and stored into various media forms such as (i) Doctrine, (ii) Tactics, Techniques and Procedures (TTPs), etc. (Nohuddin *et al*, 2010; Alkhred *et al*, 2016). One of the examples military explicit knowledge is UN Peacekeeping Training Manual or also known as 'Training Guidelines for National or Regional Training Programmes' that can be downloaded from the UN website (United Nation Website, 2016). This training manual contains several basic topics on weapon training, general military, training in UN operating techniques, safety measures and precautions, specialized training areas and exercises. For the experiment purposed, the UN Peacekeeping Training Manual will be used as datasets.

3. THE FRAMEWORK FOR TEXT ANALYTICS IN MILITARY EXPLICIT KNOWLEDGE

Figure 2 presents the framework for Text Analytics of Unstructured Data (TAUD). It consists of the three (3) main phases: (i) Data Collection; (ii) Text Data Pre-processing in Military Peacekeeping Document and (iii) Analyzing and Visualizing Selection of Terms. The first component consists of the pre-processing task while the second component is mainly focused on analyzing and visualizing the selection keywords/terms. The input materials being processed in this framework was a collection of text data. The details of each module are discussed in the following sub-sections.



```
              Terms
Docs           area cover data deploy experi forc
  chap1.txt       4     3    1      3      2    7
  chap2.txt       9     3    5      8      3   18
  chap3.txt      20     7    7     12      1    3
  chap4.txt       8     3    5      5      3    2
  chap5.txt      21    20    8     10      6   34
  chap6.txt       7     3    5      7      2    3
  chap7.txt      19     8    6      7      6    7
```

Figure 2. A Framework for Text Analytics in Military Peacekeeping Document

3.1 Document Collection

For the investigation purposes, the dataset used for this study is collected from the United Nations (UN) Peacekeeping Training Manual which can be downloaded from the UN website [4]. Basically, this sample dataset contains seven (7) main chapters such as follows: (i). the background of UN peacekeeping, (ii) the weapon training (iii) general military areas (iv) training in UN operating techniques, (v) safety measures and precautions, (vi) specialized training areas and (vii) exercises. In this phase, we converted the sample of dataset from PDF file into plain text format using the online converter. After that, we divided the chapters into seven (7) documents of plain text.

3.2 Text Pre-processing in Military Peacekeeping Document

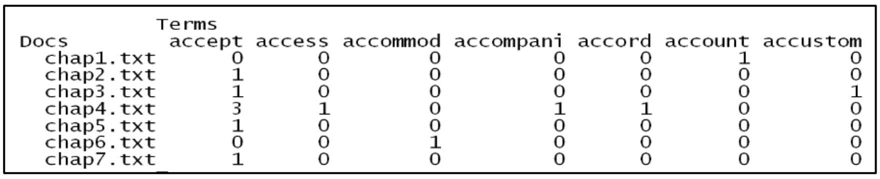
The most important step in the text analysis is the text preprocessing or sometimes known as text cleansing. Usually, the collection of raw text data tends to be dirty and contain of missing data, noisy data (containing random errors, incomplete attributes, inconsistent, etc.) and inconsistent data are common. Text cleansing is an iterative process because there are always problems that are always overlooked at the first time. According to Forman and Kirshenbaum (2008), the text data may cause to low quality of data that will affect the accuracy of mining results. Therefore, text cleansing technique is very important step as it will improve the efficiency the text mining process and provide an accurate result. Before the analysis of peacekeeping training manual can be applied, the preprocessing step is carried out to clean up the text documents into two (2) steps: (i) removing stop words and (ii) stemming. In this study, we divided the peacekeeping training manual into seven (7) text documents. In the first method of text pre-processing, all text documents are converted into lowercase whereas numbers, punctuation marks, whitespace and symbols are then removed. The purpose of converting the text documents into a standard case (lower case) is, to ensure that multiple forms of words such as ‘Training’ and ‘training’ are identified similar. The next step of text pre-processing is to remove the stop words or common keywords that do not contain important significance to be used from the text documents. These include words such as ‘is’, ‘a’, ‘an’, ‘the’, ‘it’, etc. that need to be removed in order to get accurate results. After that, stemming step is carried out to reduce the variant form of a keyword into a common grammatically root. For example, similar keywords with different spellings are grouped together, e.g., ‘training’, ‘trains’ and ‘trained’ may result the stem keyword ‘train’. In this experiment, we also remove a list stop words (see Table 1) that are found redundant and irrelevant semantically before the mining process.

Table 1. List of infrequent words

relate, name, whose, made, etc., a, b, c, certain, eg. Name, like, particular, one, brief, similar, n, r, s, decript

3.3 Analyzing and Visualizing the Military Peacekeeping Document

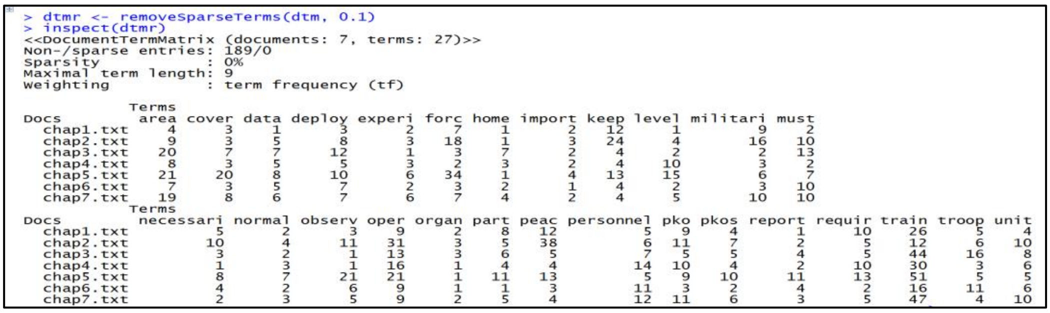
In this phase, all cleaned text data is then ready for text analytics and visualizing them into various forms: Document Term Matrix (DTM), frequency graph, word cloud and cluster dendogram. Before the graph can be plotted, we need to generate the DTM. As illustrated in Figure 3, DTM is a matrix that consists of the occurrences of words or terms in documents. In DTM, if the word exists in a particular document, the matrix entry corresponding to that row and column is presented as 1, otherwise it will be recorded as 0. Let say, if the word appears triple in that specific document then it will be recorded as three (3) in that particular matrix entry.



```
               Terms
Docs            accept access accommod accompani accord account accustom
  chap1.txt          0      0        0         0      0       1        0
  chap2.txt          1      0        0         0      0       0        0
  chap3.txt          1      0        0         0      0       0        1
  chap4.txt          3      1        0         1      1       0        0
  chap5.txt          1      0        0         0      0       0        0
  chap6.txt          0      0        1         0      0       0        0
  chap7.txt          1      0        0         0      0       0        0
```

Figure 3. A screenshot of Document Term Matrix (DTM)

DTM presented in Figure 3 consists of 1444 terms in 7 documents with 72% of sparsity. What we can observe here is that, this DTM is invariably sparse where 72% of the row entries are recorded as zero (0). As seen above, the words ‘account’, ‘access’, ‘accommod’, ‘accompani’, ‘accord’, and ‘accustom’ only appear once in a few documents. R provides a support function to remove the sparse terms in tm package. For example, in this experiment, we set 10% of empty space using the following command line: *dtmr <- removeSparseTerms(dtm, 0.1)* where the result is then stored in variable *dtmr*. Figure 4 shows the results of removing sparse in DTM with 27 terms in 7 documents with 0% of sparsity.



```
> dtmr <- removeSparseTerms(dtm, 0.1)
> inspect(dtmr)
<<DocumentTermMatrix (documents: 7, terms: 27)>>
Non-/sparse entries: 189/0
Sparsity           : 0%
Maximal term length: 9
Weighting          : term frequency (tf)

          Terms
Docs        area cover data deploy experi forc home import keep level militari must
  chap1.txt    4     3    1      3      2    7    1      2   12     1        9    2
  chap2.txt    9     3    5      8      3   18    1      3   24     4       16   10
  chap3.txt   20     7    7     12      1    3    7      2    4     2        2   13
  chap4.txt    8     3    5      5      3    2    3      2    4    10        3    2
  chap5.txt   21    20    8     10      6   34    1      4   13    15        6    7
  chap6.txt    7     3    5      7      2    3    2      1    4     2        3   10
  chap7.txt   19     8    6      7      6    7    4      2    4     5       10   10
          Terms
Docs        necessari normal observ oper organ part peac personnel pko pkos report requir train troop unit
  chap1.txt         5      2      3    9     2    8   12         5   9    4      1     10    26     5    4
  chap2.txt        10      4     11   31     3    5   38         6  11    7      2      5    12     6   10
  chap3.txt         3      2      1   13     3    6    5         7   5    5      4      5    44    16    8
  chap4.txt         1      3      1   16     1    4    4        14  10    4      2     10    30     3    6
  chap5.txt         8      7     21   21     1   11   13         5   9   10     11     13    51     5    5
  chap6.txt         4      2      6    9     1    1    3        11   3    2      4      2    16    11    6
  chap7.txt         2      3      5    9     2    5    4        12  11    6      3      5    47     4   10
```

Figure 4. An example of screenshot for DTM after removing sparse terms.

As presented in Figure 4, the six (6) most frequent terms in seven (7) documents are “train” (226) followed by “oper” (108) , “area” (88), “peac” (79), “forc” (74) and ‘”keep” (65). The analysis process of military peacekeeping document is then assisted by having visualization of term frequency graph, word cloud and cluster dendogram.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the experimental results for the proposed framework of TAUD are presented and discussed. The preprocessed dataset consists of a DTM of seven (7) documents and 27 terms were used as an input dataset for analyzing the military peacekeeping documents. In this study, we applied the Term Frequency - Inverse Document Frequency (TF-IDF) as the term weighting scheme. Table 2 shows the 27 terms from the seven documents.

Table 2. 27 terms from seven text document

Area, cover, data, deploy, experi, forc, home, import, keep, level, militari, must, necessari, normal, observ, oper, organ, part, peac, personnel, pko, pkos, report, requir, train, troop, unit

The next step is to plot a basic frequency graph where the condition of terms should occur more than 40 times using the *ggplot2* package. As illustrated in Figure 5, three (3) of the 27 terms, “train”, “oper” and “area” had the highest frequencies in the seven (7) documents which are clearly shown in the graph. This indicates that these terms are frequently used words in those 7 documents.

Figure 5 shows a word cloud which highlights the most frequently used words in text documents. Word cloud is easy to understand and readable as it provides wide choices of colors that symbolizing the words with different sizes. As illustrated in Figure 5, the keyword “train”, “oper”, “peac” and “area” are the top four (4) most important terms which validates that seven (7) documents present information that related to peacekeeping operation and training. In this experiment, we customize the word cloud as follows: the maximum number of keywords to be plotted is set to 100. It indicates that the size of keywords is corresponding to the frequency of the keyword.

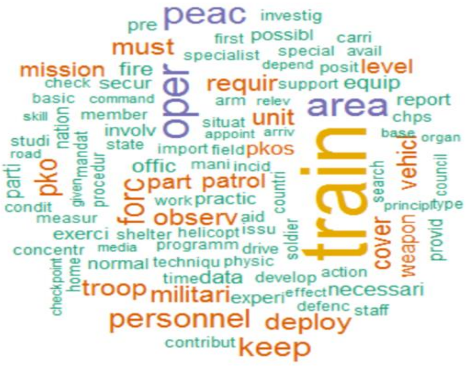


Figure 5. Word cloud with 100 most frequently occurring words

Figure 6 presents the cluster dendogram for the 27 terms in the seven (7) documents. From the dendrogram, we can see that the cluster analysis has placed “train” in the first group; “home”, “import”, “organ”, “experi”, “normal”, “data”, “report”, “necessary”, “part”, “pkos” in the second group; the “troop”, “unit”, “deploy”, “must”, “personnel”, “pko”, “require” in the third group; “keep”, “military”, “oper”, “peac” in the fourth group; and “area”, “forc”, “observ”, “cover”, “level” in the fifth group.

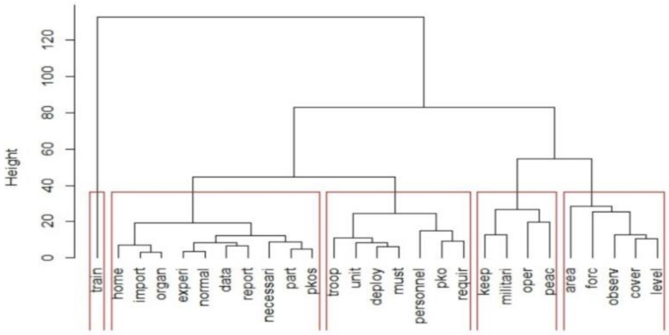


Figure 6. Cluster Dendogram

5. CONCLUSIONS AND FUTURE WORK

This paper discusses the technique of analyzing and visualizing the unstructured text data in discovering the useful knowledge from military peacekeeping document The framework for this is proposed which comprises of three main stages: (i) Data Collection; (ii) Text Data Pre-processing in Military Peacekeeping Document and (iii) Analyzing and Visualizing Selection of Terms. Based on the experiments, 27 terms in 7 documents were used as an input dataset for analyzing the military peacekeeping documents. For future work, this research will extend the dataset to include all tacit and explicit military peacekeeping documents that can give more understandings for discovering useful knowledge.

ACKNOWLEDGMENT

The authors would like to thank Universiti Pertahanan Nasional Malaysia (UPNM) and Kementerian Pendidikan Malaysia (KPM) under NRGS/2013/UPNM/PK/P3 for sponsoring this publication.

REFERENCES

Alkhred, F., Nohuddin, P.N.E & Zainol, Z. (2016). Sharing Explicit Knowledge: Designing a Peackeeping Operation Databank. In *Computational Science and Technology (ICCST),* 2016 International Conference.

Doan, A., Naughton, J. F., Ramakrishnan, R., Baid, A., Chai, X., Chen, F., & Gokhale, C. (2009). Information extraction challenges in managing unstructured data. ACM SIGMOD Record, 37(4), 14-20. Forman, G., & Kirshenbaum, E. (2008). Extremely fast text feature extraction for classification and indexing. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1221-1230). ACM.

Feldman, R & Dagan, I. (1995) Knowledge Discovery in Textual Databases. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95).* pp. 112-117.

Karanikas, H., Koundourakis, G., Kopanakis, I. and Mavroudakis, T. (2009). A Temporal Text Mining Application in Competitive Intelligence. In *Proc. Data Mining and Knowledge Discovery Stream in 23rd European Conference on Operational Research (EURO 2009).*

Massey, G. (2015) In context: Extracting Relevance from Unstructured Medical Data. *Patient Safety & Quality Health.* http://www.psqh.com/analysis/in-context-extracting-relevance-from-unstructured-medical-data/

Nohuddin, P. N. E., Ismail, R. A., & Isa, M. R. (2010). Knowledge management in military: A review for Malaysian Armed Forces communities of practices. In the *Roles of the Humanities and Social Sciences in Engineering (ICoHSE),* 2010 2nd International Conference.

United Nations Peacekeeping Training Manual. Available: http://www.usaraf.army.mil/documents_pdf/READING_ROOM/UNpeacekeepingTngMan.pdf [Accessed online 13 Mar 2016]