A REVIEW OF TEXT MINING AND KNOWLEDGE DISCOVERY IN UNSTRUCTURED TEXT ANALYSIS

[1]KALOMA USMAN MAJIKUMNA, [2]MUSTAFA ULAS, [3]UMIT FERIT ALDIM

[1,2]Department of Software Engineering Turkey,[3] The School of Foreign Languages Turkey
Email: [1]kalomausman@gmail.com, [2]mustafaulas@gmail.com, [3]silentoption12@gmail.com

Abstract: Text mining (TM) is the process of discovering hidden meaning or information that is not known previously from unstructured text. Research interest in TM has been increasing due to the availability of huge amount of data on the internet. Researchers proposed several text mining techniques that help to discover new information from unstructured text. Some of the proposed text mining techniques include: NLP techniques, Naïve Bayes, and SVM. The TM techniques help in Sentiment Analysis, Text Summarization, Opinion Leaders, and Trends. There is need of gathering and analyzing the existing TM techniques so that researchers will find it easy to learn about the previous, current and future works. In this paper, we analyzed the existing text mining techniques, its challenges, limitations, future works and implemented a text classification with Naïve Bayes.

Keywords: Text Mining Techniques, SVM, Naïve Bayes, Sentiment Analysis, Opinion Leaders, Text Summarization.

I. INTRODUCTION

Text Mining (TM) also refers to as text data mining, is the process of discovering hidden meaning or information that is not known previously from unstructured text. [1] TM research field has become very popular recently because each and every day there is increased in the availability of unstructured text data on the internet. People express their feelings on social media websites such as Facebook, Twitter, and Google+ etc. Comments on the social media sites is one of the sources of data for mining and extracting new information. [2]

TM researchers proposed several techniques for discovering new information from unstructured text that is not known by anyone previously, TM techniques helps in Sentiment Analysis, Clustering, Text Summarization, Opinion Leaders, and Trends Discovery. [1]

TM gets lots of it features from Machine Learning and Data Mining, for example algorithms for clustering, classification and text Summarization like Support Vector Machine (SVM), Neural Networks, Cross Validation and Naïve Bayes are from machine learning. [1]

Among the important issues in TM is sentiment analysis, the tasks in sentiment analysis is usually to classify given text as positive, negative or neutral. Beside classifying topics into positive, negative or neutral another task in TM is classifying a given text into predefined classes, for example SVM or Naïve Bayes can be used to classify a giving documents or text into class A or B with the help of prepared training data.

Now, let's look at some of the applications or importance of TM in our daily lives. TM can help Businesses and corporation to get feedback from their target consumers in order to know exactly how to improve the quality of their products or services. [3] Previous methods of getting user or customer feedback is questionnaire, there is no doubt that it is easier to collect and analyze user or customer feedback on social media than using questionnaire. Sentiment analysis of people's comment on social media site such as Twitter or Facebook can easily clarify if consumers are satisfied with a products or not.

Expressing opinion and sentiment in written form on social media is very common among individuals, organizations, businesses, consumers and celebrities. The manner in which opinion leaders express their feelings or opinions on a particular topic has great effect on peoples' life, such as decision making in politics, quality of product, reliability of information and positive or negative impression of event etc. [4]

II. EXTRACTING KNOWLEDGE WITH TEXT MINING

There are certain steps that are used in order to transform the unstructured text to a format that is suitable for computer processing.

2.1. Text Mining Steps

The steps in text mining as described in [5] and [6] it starts with Documents Collection, usually there are special tools for text collection such as Twitter Api and Zapier. Then text pre-processing will be applied, text pre-processing includes stop word removal, http address removal etc. Next step is the analysis of the text, it includes text clustering, classification and summarization. The resulting information retrieved can be place in management information system and finally the knowledge will be extracted. Figure 1 shows the mining steps.

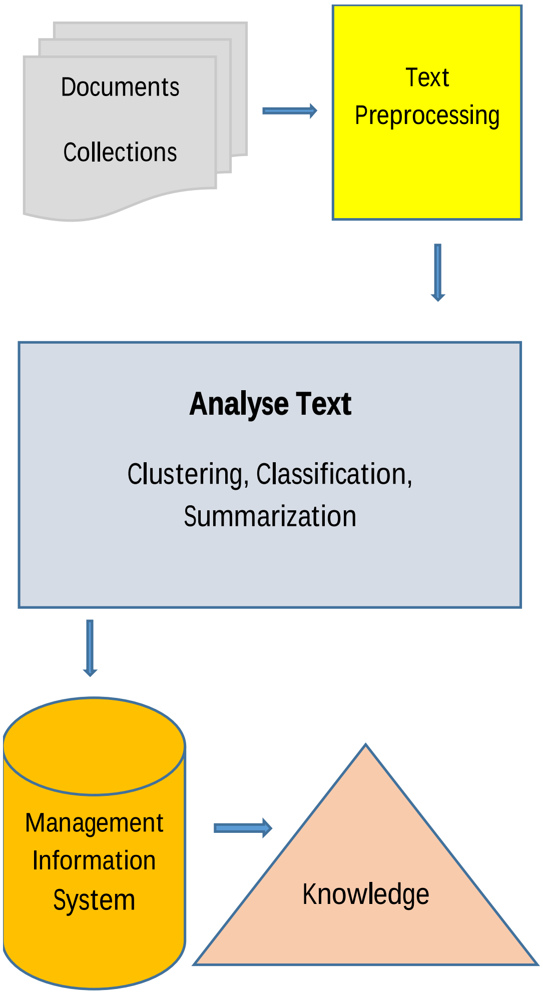

Figure 1 Mining steps

III. APPLICATION FIELD OF TEXT MINING

The purpose of applying text mining differs based on the users' need. As described by Gupta and Lehal [5] most of the purpose of text mining are as follows:

- Categorization (or Classification)
- Clustering
- Summarization
- Feature Extraction.
- Text-based navigation.
- Search and Retrieval [5]

In addition to the above list proposed by Gupta and Lehal sub topics were added i.e. Trends and Opinion Leaders. Due to space constraints on this article discussion were only made about Classification, Clustering, Summarization, and Opinion Leaders.

a. Classification

The goal of classification is to separate or classify document into two or more groups, for example classify text as economy news or sport news; classify sentiment as negative or positive. The most well known in classification is sentiment analysis, the task in sentiment analysis is usually classifying text into Positive Negative or Neutral. [7] Classification is based on supervised learning with predefined training data. The most used algorithm for text classification include: K-nearest neighbor, Neural Networks, Maximum Entropy, Naïve Bayes, and Support Vector Machine (SVM). [7] According to findings by Akaichi, Dhouioui and Pérez in [2] SVM outperforms the other classification methods.

b. Clustering

Clustering as stated by Kunwar in [8] is the most common and simple unsupervised learning problem and it is defined as “the process of organizing objects into groups whose members are similar in some way”. Clustering plays similar role as classification problem, however clustering is unsupervised learning problem whereas classification is supervised learning problem. [9]

c. Text Summarization

Text summarization is an important problem nowadays due to the availability of large text data on the internet. As stated byRadev, Hovy and McKeown in [10]text summarization refers to a text generated from single or multiple text source, which conveys essential information that is in the original document(s), and which is usually shorter than half of the original text(s).

The two different types of text summarization are Single Document and Multiple Documents Summarizations. The most common method for text summarizaton has been described in [11] as follows:

- Single Document Summarization methods includes:
  - Naive-Bayes Methods
  - Rich Features and Decision Trees
  - Hidden Markov Models
  - Log-Linear Models
  - Neural Networks and Third Party Features
  - Deep Natural Language Analysis Methods
- Multiple Documents Summarizaton methods includes:
  - Abstraction and Information Fusion
  - Topic-driven Summarization and MMR
  - Graph Spreading Activation
  - Centroid-based Summarization

Details about text summarization can be found in [11][12] [13]

d. Opinion Leaders

Opinion leader has been defined by study.com as a well-known individual or organization that has the ability to influence public opinion on the subject matter for which the opinion leader known. Opinion leaders can be politicians, business leaders, community leaders, journalists, educators, celebrities and sports stars. [14]

There are three important entities in opinion leaders, these entities are (a) the target entity on which opinion was expressed, (b) an author or the opinion leader who expressed the opinion, and (c) feeling or sentiment about the entity held by the opinion leader or the author. [15]

Most of the opinion leader detection systems uses sentiment analysis techniques. As far as we found in literature some of the most detailed work on opinion leader is proposed by [16] [17] [18] for further reading.

IV. CHALLENGES AND FUTURE WORKS IN TEXT MINING

Most of the works if not all, in the TM literature concentrated on developing language dependent techniques with the majority of the works in English language. As in [7] [2] [19]and many other researches the NLP techniques were dependent on English language, except Stanford University NLP group that extends it work to Arabic, Chinese, and German text. [20] So, multilingual mining techniques is a challenging problem and could serve as future research field.

Another open area in TM research is finding Opinion Leaders, Binali, Potdar and Wu in [3] found that current research on opinion leader concentrated on products reviews on a single item, but very little research is done on explicit item and it comparison with other items which can be very beneficial to consumers and producers.

V. APPLICATION OF TEXT CLASSIFICATION WITH NAIVE BAYES

The general idea behind text classification have been discussed in the previous section, this section aims to describe one example of text classification that we did in order to give the reader a better understanding of classification process. The methods and procedures of classification of the text (Turkish Language Text) into two classes (economy news or sport news) was presented with detail explanation of the approach.

Naive Bayes with Java Programming was used to classify news written in Turkish Language, the news samples were sport and economy news as follows.

$$\mathbf{P}(\boldsymbol{sport}|\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn}) = \frac{\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn}|\boldsymbol{sport})\boldsymbol{P}(\boldsymbol{sport})}{\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn})}$$

$$<>$$

$$\boldsymbol{P}(\boldsymbol{econ}|\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn}) = \frac{\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn}|\boldsymbol{econ})\boldsymbol{P}(\boldsymbol{econ})}{\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn})}$$

$where$ $\boldsymbol{x1},\boldsymbol{x2},\dots.\boldsymbol{xn}$are the words that appear in the texts samples.

Because $\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn})$is common in both classes we eliminate it and our equation becomes

$$\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots\boldsymbol{xn}|\boldsymbol{sport})\boldsymbol{P}(\boldsymbol{sport})$$

$$<>$$

$$\boldsymbol{P}(\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\dots|\boldsymbol{econ})\boldsymbol{P}(\boldsymbol{econ})$$

After expansion the equation will be equal to the below equation

$$\boldsymbol{P}(x_1\,|\boldsymbol{sport})P(x_2\,|\boldsymbol{sport})P(x_3\,|\boldsymbol{sport})\dots P(x_n\,|\boldsymbol{sport})$$

$$<>$$

$$P(x_1\,|\boldsymbol{econ})P(x_2\,|\boldsymbol{econ})P(x_3\,|\boldsymbol{econ})\dots P(x_n\,|\boldsymbol{econ})$$

If for example xi=0 in one of the equation, then everything becomes zero like $P(C)\prod_{i=1}^{N}P\left(x_i\,\middle|C\right)$ = 0, so to avoid 0 problems that will make everything zero log was taken as described below.

log(P(x1/ $\boldsymbol{spor}$t)) + log(P(x2/ $\boldsymbol{sport}$)) + …+ log(P(xn/ $\boldsymbol{sport}$))

<>

log(P(x1/ $\boldsymbol{econ}$)) + log(P(x2/ $\boldsymbol{econ}$)) + …+ log(P(xn/ $\boldsymbol{econ}$))

The texts were classified as economy news or sport news based on the class with higher probability.

a. Approach

Text samples of 35 sport news and 35 economy news were collected from various online newspapers that were written by various authors. The words that appeared inside the sport and economy texts samples served as the training set that the program learned from.

b. Algorithmic Approach

i. An alternative of Hap Map was used to store the list of words after reading from files, i.e. Multiset and Hash Multiset classes of Guava library of Google, Guava library of Google was selected because it is easy to implement it with Java Programming.

ii. While calculating the probability logarithms of the result were added instead of multiplying the probability of the result so that we get rid of 0 probabilities that will make the result 0 as described with mathematical terms above. But taking logarithms in Java create another problem, because log 0 is negative infinity (-∞) in Java. To get rid of negative infinity log() function was re-written so that it returns large negative value instead of negative infinity (-∞).

iii. Stemming was applied to all words in order to get the root words, so first 3, 4, 5 or 6 characters of all words were selected as the root words. Example of stemming in Turkish language text (Geliyorum, Geliyorsun, Geliyor, Geliyoruz, Geliyorsunuz, Geliyorlar. All these six words can be consider as “Gel” because the root word for all of the words is the same).

iv. All the words were converted to lower case letters.

c. Result

Finally, this program shows outstanding performance because more than 150 different files of sport and economy news were tested and it gives performance of more than 90% efficiency.

CONCLUSION

This paper discussed the current works on Text Mining (TM), it is challenges, possible future works and implemented a text classification with Naïve Bayes. Text Mining also refers to as text data mining, is the process of discovering hidden meaning or information that is not known previously from unstructured text. [1] TM is one of the research fields that attracts researchers globally to work on it due to the daily increased in the availability of large amount of text on the internet. The basic steps in TM approach has been examined. The application fields of text mining that ranges from Classification, Clustering, Text Summarization and Opinion Leaders has been discussed.

REFERENCE

[1] D. S´anchez, M. Mart´ın-Bautista and I. Blanco, "Text Knowledge Mining: An Alternative to Text Data Mining," *2008 IEEE International Conference on Data Mining Workshops*, p. 1, 2008.

[2] J. Akaichi, Z. Dhouioui and M. J. L.-H. Pérez, "Text Mining Facebook Status Updates for Sentiment Classification," *IEEE*, 2013.

[3] H. Binali, V. Potdar and C. Wu, "A State Of The Art Opinion Mining And Its Application Domains," *IEEE*.

[4] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini and F. Menczer., "Predicting the Political Alignment of Twitter Users," *IEEE*, 2011.

[5] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1*, 2009.

[6] G. Shi and Y. Kong, "Advances in Theories and Applications of Text Mining," *The 1st International Conference on Information Science and Engineering (ICISE2009)*, 2009.

[7] M. Farhadloo and E. Rolland, "Multi-Class Sentiment Analysis with Clustering and Score Representation," *IEEE 13th International Conference on Data Mining Workshops*, 2013.

[8] S. Kunwar, "codeproject.com," 2013. [Online]. Available: http://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm. [Accessed 2015].

[9] J. D. Martin, "Clustering Full Text Documents," *IJCAI-95 Workshop on Data Engineering for Inductive Learning*, 1995.

[10] D. R. Radev, E. Hovy and K. McKeown, "Introduction to the special issue on summarization. Computational Linguistics.," 2002.

[11] D. Das and A. F. Martins, "A Survey on Automatic Text Summarization," Language Technologies Institute Carnegie Mellon University, 2007.

[12] E. Lloret, "TEXT SUMMARIZATION: AN OVERVIEW," Universidad de Alicante Alicante, Spain, 2006.

[13] E. I. Mani and M. Maybury, "Automatic summarising: factors and directions," *The MIT Press*, 1999.

[14] "Opinion-Leader in Marketing," http://study.com/academy/lesson/opinion-leader-in-marketing-definition-lesson-quiz.html, 2015.

[15] R. Arora and S. Srinivasa, "A Faceted Characterization of the Opinion Mining Lanscape," *IEEE*, 2014.

[16] B. Liu, M. Hu and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *WWW '05 Proceedings of the 14th international conference on World Wide* ,2005.

[17] P. Sobkowicz, M. Kaschesky and G. Bouchard, "Opinion mining in social media," *Science Direct*, 2012.

[18] K. Khan, B. Baharudin, A. Khan and A. Ullah, "Mining opinion components from unstructured reviews," *Science Direct*, 2014.

[19] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques," *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, 2003.

[20] "The Stanford NLP Group." The Stanford NLP (Natural Language Processing) Group," [Online]. Available: http://nlp.stanford.edu/. [Accessed 2 May 2015].

★★★