Mehran University Research Journal of Engineering & Technology
Vol. 38, No. 4, 901-914 October 2019
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1904.04

Bringing Shape to Textual Data – A Feasible Demonstration

ANOUD SHAIKH*, NAEEM AHMED MAHOTO*, AND MUKHTIAR ALI UNAR*

RECEIVED ON 07.05.2018 ACCEPTED ON 30.10.2018

ABSTRACT

The Internet has revolutionized the communication paradigm. This has led towards immense amount of unstructured data (i.e. textual data), which is a major source to get useful knowledge about people in several application domains. TM (Text Mining) extracts high quality information to discover knowledge by drawing patterns and relationships in textual data. This field has taken great attention of the research community. As a result, several attempts have been made to propose, introduce and refine techniques applied for uncovering knowledge from text data. This study aims at: (1) presenting existing TM techniques in the scientific literature, (2) reporting challenges/issues and gaps that still need attention, and (3) proposing a framework to bring shape to textual data. A prototype has been developed to demonstrate the effectiveness and potential worth of proposed approach to display how unstructured data (i.e. news articles in this study) has been brought to a shape representing interesting knowledge. The proposed framework implements basic NLP (Natural Language Processing) functions in combination of AYLIEN API (Application Programming Interface) functions. The results reveal the fact that how events, celebrities and popular news-items have been covered in the electronic media, and it also represents subjectivity of topical news events. The news coverage trends highlight the significance of daily news events, which may assist in getting insight about the media groups.

Key Words: Text Mining, Text Analytics, Knowledge Discovery, Unstructured Data, Visual Representation.

1. INTRODUCTION

The modern technological growth has increased tremendous amount of data and every day the amount of data is growing at an increasingly high speed. The number of tweets on Twitter, for example, has risen to 6,000 tweets per second on average, which corresponds to approximately 500 million tweets per day. Likewise, Facebook scans 105 TB of data every 30 minutes. In 2017 alone, it was reported that there were 5.2 billion daily Google searches and 269 billion emails sent [1]. Literature also represents a large portion of textual data in the form of books, journals, and theses. Thus, every second of every minute of every day, new data is born. It has been reported that 90% of the data has been created in the last two years alone [2-3]. According to Ben Walker 2.5 quintillion bytes of data are created every day, which would require 10 million blue-ray discs that if stacked on one another would measure the height of 4 Eiffel towers [4]. People, businesses, and devices have

Authors E-Mail: (anoudmajid85@gmail.com, naeem.mahoto@faculty.muet.edu.pk, mukhtiar.unar@faculty.muet.edu.pk)
* Institute of Information & Communication Technologies, Mehran University of Engineering & Technology, Jamshoro, Pakistan.

all become data generators that are pumping out enormous amounts of data to the web each day. Undoubtedly, there are lots of insights within this data that cannot be ignored. It is important to bring structure to this text, which has lot of possibilities and insights to look for relationships, hierarchies, patterns and trends to discover knowledge. Manually exploring and analyzing such a large collection of data in order to find new insights for prediction and forecasting is unrealistic. Also, this data is beyond the capabilities of traditional applications for exploratory analysis. High volumes of data need to be analyzed for the discovery of trends and meaningful patterns, which are vital for effective decision-making. There are several tools and techniques for mining the text data effectively and extracting the new outcomes and the rich insights it can bring. The TM techniques that aim at providing such insights include text categorization, clustering, summarization, concept extraction, topic detection, information retrieval and prediction.

The paper aims at: (1) To present a comprehensive study about existing methods and techniques in the TM, (2) To determine the challenges, strengths and gaps (limitations), and (3) To propose a framework to bring shape to textual data. Though many efforts have been already made in this domain, this study highlights the recent research methods applied in TM. Besides, the advanced techniques have not been adopted in proposed framework to discover knowledge with lightweight functions. The demonstration of news trends in visual formats validates the effectiveness of the proposed framework.

The organization of the paper is as follows. Section II discusses lifecycle of TM; section III reports roots of the TM. The text mining techniques are presented in section IV, whilst section V addresses TM applications. The challenges and issues in TM are reported in section VI. The proposed framework is described in VII and conclusions are drawn in section VIII.

2. LIFECYCLE OF TEXT MINING

Text Mining: TM or TDM (Text Data Mining) is a process of discovering interesting and actionable information to look for emerging trends and patterns, which are valid, useful, unexpected and understandable [5]. This means through TM novel knowledge is captured automatically, which provides new and exciting information about the world [6]. TM is considered to be a specialist technology requiring a multitude of skills such as linguistic background, statistics, computational skills, and psychology [6]. It enquires tools and techniques that vary in usability, accessibility and configurability as these enable a deeper analysis of the information. Besides, it helps in understanding and identifying business insights in the content and also highlights relationships between texts in a document or corpus, which would otherwise be undiscovered and ignored. The TM in the business domain is referred to as text analytics [7].

TM has potential to enrich information and knowledge management processes. It can explore large amounts of text containing extensive and detailed coverage of innumerable observations. To take advantage from the text, it is necessary to bring some level of structure into textual data format such that most of the available clues are identified and necessary actions may be taken on timely basis. Although, computer program is linguistically challenged but the current processing speed [8] has adequate power and potential to make and benefit from efficient tools, techniques and algorithms.

Lifecycle of Text Mining: It contains eight steps, which are described in Fig. 1.

Step-1: Problem Definition and Defining Specific Goals: This step identifies the fundamental problem that needs to be solved. Thus, it enables to determine the right content needed for mining and the right TDM approach. Understanding of the problem definition is considered the key to carry out research in TM. A good knowledge of the problem and accurately defining it builds a path to reaching towards a solution or a recommendation. TDM analyzes the contents of data as well as it also explores the outcomes so that new connections or patterns are detected that are manually next to impossible.

Step-2: Use/Design and Build/Outsource TDM Approach: A ready-made TDM approach is chosen, only if the problem investigation person has good technological skills. In case, investigator possesses good programming skills, then it is better to build a TDM approach. There are possibilities that the approach may be outsourced as solving a TM problem that may require good NLP knowledge.

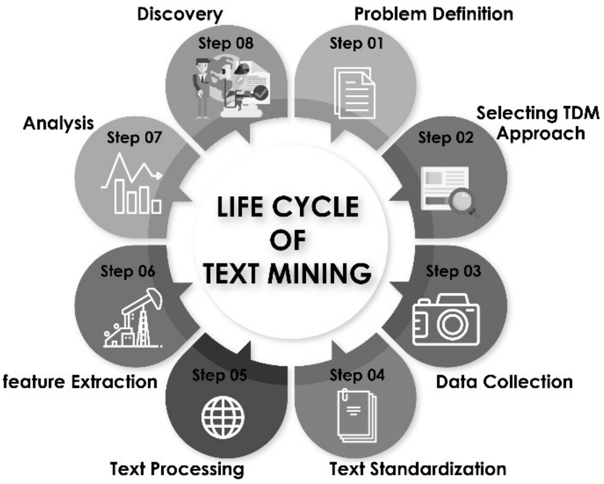

FIG. 1. LIFE CYCLE OF TEXT MINING

Step-3: Determination of Text – Text Data Collection: In certain scenarios, relevant documents for mining may already be available; whilst in other cases, data need to be drawn from the web manually or through web crawlers. The key challenge lies at deciding the duration of the data, which has to be collected. On one hand, large data collection increases the cost and complexity; on the other hand, it also leads towards better accuracy. Therefore, it is necessary to keep certain factors while collecting the dataset such as length, duration, and cost. The collected set of documents, often, comprises of large amount that may require data sampling techniques for selecting a set of relevant and/or targeted documents.

Step-4: Text Standardization: The collected documents could be in a variety of forms, which may be transformed into suitable format for further processing. For instance, documents can be transformed into XML (Extended Markup Language) as a standard exchange format. In the simplest terms, XML is a standard way to insert tags onto a text and identify its parts. The transformation process may prune irrelevant information such as menus, adverts, copyright information, and templates.

Step-5: Text Preprocessing: The set of documents, once available in standard format, has to be analyzed using several components of NLP. These components are tokenization, filtering or stop-words removal, stemming or lemmatization and normalization. **Tokenization** breaks the stream of characters into words/tokens. The characters like space, tab and newline are delimiters and they do not fall into category of tokens. Later, the tokens are filtered, since stopwords do not contain relevant information; this component is termed as **stopwords removal**. The remaining tokens are converted into a standard representation known as stemming or lemmatization. **Stemming** (morphological analysis in linguistics) regularizes grammatical variants, for instance, singular/plural or past/present. Finally, **normalization**

converts each token into either upper case or lower-case. Spelling normalization is also done depending on the set of documents. The normalized tokens or words are then represented in several different models; for example, BOW (Bag of Words), Vector Space Model, Document Matrix. Each of these models helps in further processing.

Step-6: Feature Extraction/Build Concept and Category Model: The text features could be identified on three levels, which are words, sentences and documents [9]. The features are selected from, for instance, representation model present in the considered data collection [10]. Feature selection is an important step, since it reduces the complexity by selecting essential features and ignoring irrelevant ones. Feature Extraction also helps in reducing the dimensionality.

Step-7: Analysis: In this step, DM (Data Mining) methods are applied to find out relationships or patterns between tokens. The relevant facts and relationships are determined and are extracted in a structured form for faster review and analysis. The facts are connected in new ways to synthesize knowledge for creating actionable insights. There are different algorithms, which can be chosen and applied on the concepts for detailed analysis. The best-scored concepts could be merged with other data to predict future behavior.

Step-8: Reach an Insight/Outcome/Recommendation: The selection, interpretation, evaluation and visualization of knowledge are performed at this stage [11]. There are a lot of visualization tools available and graphs or charts could be drawn for in depth understanding and better analysis to find new outcomes and exciting discovery.

3. ROOTS OF TEXT MINING

TM is a subfield of DM, which itself has grown from its parent disciplines – ML (Machine Learning), databases, data warehousing and knowledge discovery [6]. TM, being an interdisciplinary field, employs many computational technologies such as ML, NLP, AI (Artificial Intelligence), IT (Information Technology), CL (Computer Linguistics), Biostatistics, Pattern Recognition and Psychology. TM is also closely related to IR (Information Retrieval) and IE (Information Extraction). Thus, the roots of TM are scattered over several overlapping fields that are described in the following.

3.1 Data Mining

DM serves two goals. Firstly, it identifies emerging patterns and trends termed as insight, which may help in taking actions on timely basis. This provides tremendous economic value, which is often imperative to businesses looking for competitive advantage. Secondly, it helps in prediction by building a model predicting based on certain given data [10]. This helps organizations to plan, predict and forecast and take appropriate measures and provide recommendations effectively and on time. Precisely, patterns are extracted from large computerized databases, which are previously unknown and are particularly important; whilst in TM, patterns are extracted from NLP text where input is free unstructured text.

3.3 Machine Learning

ML builds algorithms taking input data and then uses statistical analysis to predict outcomes within an acceptable range. Supervised and unsupervised are often the two main approaches in ML [12]. In supervised learning, model is trained based on training data; the predictions are computed on unseen new data. In unsupervised learning, no training is provided to model. The methods categories data sets based on similarity among the data points in a given dataset.

3.3 Natural Language Processing

NLP is a component of TM that helps the machine read and understand text by performing linguistic or grammatical analysis [13]. It needs a consistent knowledge base, such as detailed thesaurus, a lexicon of words, a data set for linguistic and grammatical rules, ontology and up-to-date entities.

3.4 Information Retrieval

An IR (Information Retrieval) system finds relevant texts from a large collection of documents and presents them to the user like most search engines. The main job of IR system is providing the right information to its users at the right time with less emphasis on processing the textual information. A basic concept of IR is to measure similarity i.e. a comparison is made between the documents in a set of large document collection, measuring how similar the documents are. Documents are presented based on the keywords, sometimes resulting in thousands of hits containing irrelevant hits as well [14].

3.5 Information Extraction

IE (Information Extraction) is a technique that extracts meaningful structured information from unstructured and/ or semi-structured data format [13]. The process also needs NLP in order to read the machine-readable documents.

4. TEXT MINING TECHNIQUES

TM offers variety of techniques such text classification, clustering, building ontology, sentiment analysis, document summarization and many more. The techniques and algorithms that have been used in the existing literature reported in Table 1. The popular techniques are briefly described in the following.

4.1 Text Classification

Text Classification and categorization methods (it is supervised learning approach) examine the text document based on the presence or absence of particular features in the text; and then assign the text documents into two or more categories. The classifier is trained with training data and known outcomes for a specific task; it, later, assigns categories to new, unseen text documents. A common example of binary classification is spam filtering. Naïve Bayes and Maximum Entropy (often called Logistic classifier), both of them are supervised classifiers and are the most popular algorithms used for text classification. The performance of text classifiers is evaluated with accuracy or precision and recall.

4.2 Text Clustering

Text clustering is an unsupervised learning approach. The similar documents are grouped together based on certain similarity attributes; it does not require prior information to group together relevant documents. Text clustering is commonly used in fraud detection applications. The most commonly used algorithms for clustering are k-Means Clustering, Expectation-Maximum algorithms and Hierarchical clustering. In TM, clustering algorithms can be used to cluster product reviews, for example, based on similar combination of words and phrases.

4.3 Building Ontology

Ontology construction is a growing research topic [15]. The two main parts of ontology are concepts and relations. Domain ontology is constructed by first selecting the concepts (concept discovery) from the data source. Relevant or similar terms are determined by TF-IDF (Term Frequency, Inverse Document Frequency) scheme. Also, synonyms are determined. The resulting terms are checked and filtered by a domain expert. Later, concepts are arranged in a hierarchal structure indicating relations between the concepts and are evaluated by an expert.

TABLE 1. THE EXISTING TEXT MINING TECHNIQUES

References	Algorithm/Technique	Description	Advantages	Limitations
Liu et. al. [15]	Domain knowledge method using Meta Map and machine learning methods using sequential classifier conditional random fields	Three models of medical concept extraction are developed and are performed on English clinical texts and then compared.	The model using the combined approach of CRF (Conditional Random fields) and MetaMap features obtains better results.	Limited to medical reports have been considered as information repository.
Krishnan et. al. [17]	Information Extraction and Categorization	From the titles of scientific articles, concepts are evaluated.	Understanding of the key contribution in the articles is achieved.	The scope is limited to scientific articles.
Aga et. al. [18]	Visualization, Force Atlas algorithm	Concepts are generated from German website in the form of a concept cloud showing the main interests of the companies.	Mostly related concepts appear adjacent. Company's attention and focus is outlined.	The concepts are limited to STW Thesaurus of Economics
Xu et. al. [19]	Association rule mining	Interesting associations among concepts that co-occur is extracted. Directed and transitive associations are represented.	The approach proposed significantly reduces the number of associations.	Only association rule support and confidence are considered.
Dinh and Tamine[20]	Vector space model, BOW + word positions	Semantic indexing and retrieval through domain concepts on biomedical documents. Concept reference scoring is done.	The proposed methods facilitates and improves biomedical information indexing and retrieval.	Extracted concepts are limited to MeSH Thesaurus.
Chin et. al. [21]	Machine learning Approach and Semantic Approach	An abstract discourse-level word sense disambiguation on word clusters is performed.	The proposed approach provides a possibility of incremental learning capability for NLP based systems	Only limited to Wordnet.
Abebe et. al. [22]	Natural Language Processing (NLP)	Domain concepts and relations are extracted from program identifiers.	Supports concept location executed in the context of bug fixing.	Program element names must be chosen carefully by the programmer.
Ajgalk et. al. [23]	Classification, Page-rank algorithm	Representation of documents is shifted from keyword based to key concepts, as keywords are sometimes ambiguous but concepts are unambiguous. The presented approach is evaluated and compared to TF-IDF keyword model.	The key concepts extracted are satisfyingly accurate and are quite convenient to be used on web as it allows online key concept extraction.	A lightweight ontology WordNet is only utilized.
Szwed[24]	Rule based approach	Concepts are extracted based on correct morphological forms in Polish text. Annotations prepared by the user are compiled into transformation rules.	The proposed approach is quite general and can be applied to texts in other languages.	The approach does not work for 3-gram translation patterns.
Yong-Bin et. al. [25]	Heuristic approach: NLP+statistics+domain specific knowledge+ inner structural patter of terms	Best key concepts are selected based on candidates by leveraging the inner structure of terms . (CFinder)	Based on some metrics, best key concept candidates are selected.	Not efficient when phrase-length is large.
Khin and Lynn[27]	Statistical and Linguistic rules	The proposed method extracts ontology concepts from multiple text of the same type using mutual information and domain frequency. The corpus comprises of financial and economic reports in Chinese Language.	Extraction of relevant ontological concepts from multiple sources.	The extracted concepts are limited to two-word phrase.
Wang et. al. [28]	Word frequency analysis, Clustering	User experience is evaluated by user feedback using text mining. It also using grounded theory.	Results show the various factors which influences user's emotions.	External Validity i.e., further analysis of original data and visual networks is required.
Prameswari et. al. [29]	Sentiment Analysis, Summarization	Online hotel reviews are mined to build the hospitality sector as an integral part of Indonesia's tourism industry.	The two text mining techniques namely, summarization and sentiment analysis are combined and exciting outcomes are observed.	Sentiment graphs of positive and negative reviews are shown only in five categories.
Jiang et. al. [30]	Multi-label text classification	An embedded model for multi-label text classification is proposed based on ELM (Extreme Learning Machine+ L21-norm minimization of the hidden layer output weights matrix.	Inherits the merits of ELM, facilitates group scarcity, and reduces complexity of the learning method. Proposed algorithm obtains superior performance.	The presented approach takes more time than original ELM.
Forman and Kirshenbaum[31]	Text Feature Extraction for Classification and Indexing	A fast method for text feature extraction is proposed that folds together Unicode conversion, forced lowercasing, word boundary detection and string hash computation. The method yields word and phrase features represented as hash integers rather than strings.	It requires less computation and less memory.	There may happen a collision between important, predictive feature and more frequent word in the Hash function.
Chintan et. al. [32]	Supervised learning - classification	Text Mining is used to identify and predict cases of child abuse in a public health institution in Netherlands in medical texts.	Both structured and unstructured data are taken into account for prediction. Real dataset has been used for experiments.	The focus remained on only child abuse. The other issues would have been assessed.
Hashimi et. al. [33]	Feature selection criterion	The study addresses several selection criteria to facilitate appropriate selection of techniques in text mining.	Two categories are reported for selecting text-mining technique. 1) Generic that covers usability, comprehensiveness and flexibility; 2) Specific deals with user interfaces, level of satisfaction.	The study has considered 130 research articles related to text mining.
Shihand et. al. [34]	Document representation for text categorization	Novel use of Siamese LSTM (Long Short-Term Memory) for document representation is explored and is used in text categorization.	Siamese LSTM network facilitates the measure of semantic distance between any two different text documents by learning distributed vector representations of documents that reflects the semantic relatedness between any pair of documents.	Experiments are only conducted on IMDP and 20Newsgroups.
Wang et. al. [35]	Topic Modeling approach	A topic modeling approach for Named Entity Disambiguation for questions in English and Chinese in community question answering has been reported.	No human annotations needed for the learning process of the model.	The training time is significantly higher for the model.

4.4 Document Summarization

Summaries are constructed from NLP text. There are two main categories of summarization techniques. The first, and more commonly used, is Extractive summarization which is a summary consisting of units of IE from the original document itself. The other is Abstractive summarization, which contains information that is synthesized, as the units of information may not necessarily appear in the original text. The document summaries may also be used for clustering/categorizing documents into specified groups or taxonomies.

4.5 Sentiment Analysis

The goal of sentiment analysis (also opinion mining) is to detect sentiments or emotions of users from their textual posts. Users all over the world use blogs and social networking sites to publish their posts representing their views/opinions [16]. Business organizations want to know the opinions of their customers and spend a huge amount of money to know and attract their customers. Opinions can be positive, negative or neutral. Most common supervised algorithms for polarity classification of opinions are SVM (Support Vector Machine) and Naïve Bayesian Classifier [9].

5. APPLICATIONS OF TEXT MINING

TM has remained in the use of government organizations, law enforcement agencies, news agencies, business intelligence and customer relationship management as well as researchers. It has helped to track, trace and understand contacts between individuals, among organizations and different ideologies. The task of identifying whether news has the same story as once told by someone a year back, for example, may not be assisted manually due to its nature, which requires error-free processing and rapid response. However, a computer can never tire or lose interest and can do such tasks in the blink of an eye. TM processes dense text to seek information that lies beneath the considered data for better decision-making, to gain indispensible business insights and to mitigate operational risks.

The prominent application domains, where TM has played key role in getting in-depth knowledge are, for instance, (i) Social Media Analysis, (ii) Email Spam Filtering, (iii) Cyber Crime Analysis, (iv) eHealth Management System, and (v) Opinion Mining. Challenges in TM.

Inherently TM faces several challenges due to dealing with natural language. The foremost common challenges are illustrated in Fig. 2. These challenges are addressed in the available literature. However, it still requires attention to tackle such gaps in the TM, which makes TM a challenging field.

Ambiguity in Text: Ambiguity is text is a major challenge. For example, one word can have several different meanings and multiple words give one general meaning.

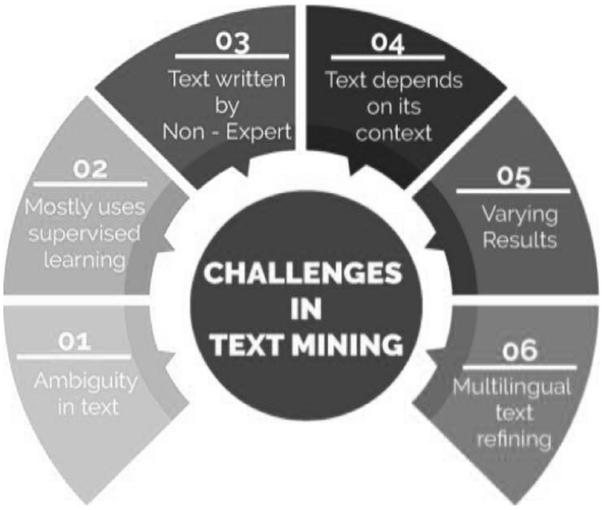
03
Text written by Non - Expert
04
Text depends on its context
02
Mostly uses supervised learning
05
Varying Results
CHALLENGES IN TEXT MINING
01
Ambiguity in text
06
Multilingual text refining

FIG. 2. CHALLENGES IN TEXT MINING

Mostly Uses Supervised Learning: Many TM techniques use supervised learning, which is useful when amount of training data is available. It is quite expensive to create training data for textual data.

Text Written by Non-Expert: The concept of a sentence, paragraph or document(s) is extracted based on the lexical choice of the writer. At times, the writer is writing fake or false text or the choice of his/her vocabulary does not convey proper meaning and therefore lacks to give audience a clear picture of the work.

Text Depends on Its Context: Most of the time, the meaning of the text could be better understood in terms of the contextual information. Thus, TM approaches need to be further extended to incorporate the context of the text for powerful text analysis.

Varying Results: Results and depth of analysis can vary wildly from vendor to vendor [36]. The outcomes of one technique are quite different in comparison with other techniques, since TM is data-driven approach and each technique has its pros and cons.

Multilingual Text Refining: Most of the times, the text to be mined needs permission from the organization and/ or author as it holds copyright legislations. This creates an issue for processing such type of data.

6. THE PROPOSED FRAMEWORK

This section reports the proposed framework to bring shape to text data. The phases involved in the framework are described in the following subsection.

6.1 The Building Blocks of Proposed Framework

The proposed framework comprised of four blocks (i.e. components) in order to shape the textual data into understandable visual information as shown in Fig. 3.

Block-1: Web Crawling: A web crawler has been developed to collect the inquired textual data form the online repositories. The news that is online available has been targeted. The collected dataset comprises of news articles and daily news published online. The daily news and news articles authored by several columnist and journalists have been stored into documents to set a document database (c.f. Definition-1). The document database comprised of 6-months daily news and news articles.

Definition-1: Document Database: Let R be a repository of collected news articles and news events. The DD (Document Database) associated with R is set of news items NT_i, such that for each $r_j \in NT$. Each entry in $DD_i = \{c_1,c_2,c_2,\ldots,c_n\}$ is a set of news items related to r_j.

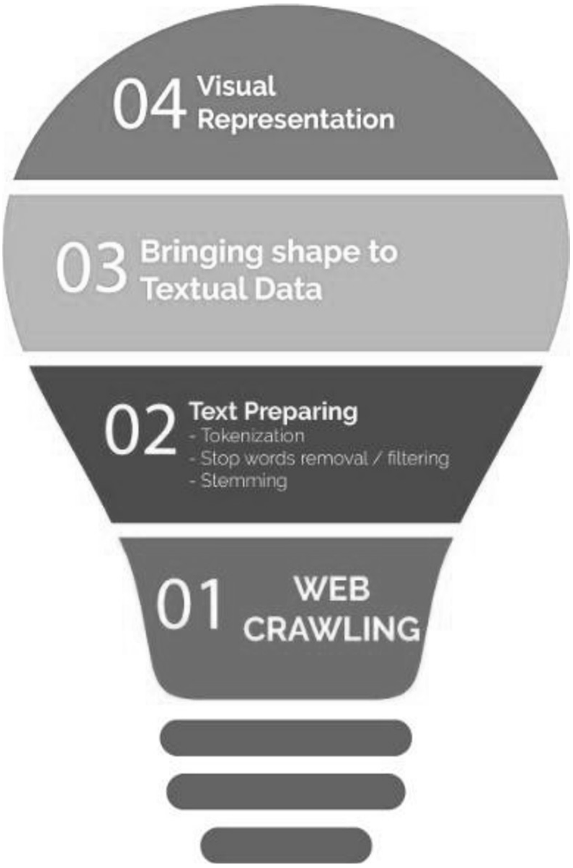

FIG. 3. BUILDING BLOCKS OF PROPOSED FRAMEWORK

Block-2: Text Preprocessing: Text processing is the major phase in TM, since neat and clean data (i.e.quality data) produces better outcomes. The collected DD has been preprocessed using necessary procedures such as tokenization, stopwords removal, and stemming.

Tokenization: Tokenization breaks the sentences into its parts (i.e. words and phrases). For instance, assume a sentence 'This study proposes a framework – Bringing shape to textual data.'; the tokenization process the sentence and results into words: {'This', 'study', 'proposes, 'a, 'framework, '-, 'Bringing, 'shape', 'to', 'textual', 'data', '.'}.

Stopwords Removal/Filtering: Stopwords, such as is, at, a, on, of etc. does not contain meaningful information that needs to be processed for knowledge discovery. Therefore, filtering is applied to prune such words from entire set of documents.

Stemming: Stemming determines base of the words. The filtered words are stemmed to trace the base of each word available in the set of documents. For instance, bring is the base word of bringing in the above-mentioned example.

The text-preprocessing phase turns entire set of documents into a suitable format ready for further processing. In order to apply TM techniques, the words are transformed into Representation Model. This study applies basic NLP functions to present the set of documents in the form of BOS.

Block-3: Bringing Shape to Textual Data: The structural information has been extracted from the processed DD with the help of pattern mining technique. The extracted patterns carry useful information.

The popular AYLIEN Text Analysis API has been used to extract the hidden useful information from the considered DD. Although several advanced NLP functions exist, but this study adopt very basic functions in the proposed framework. In particular, the patterns are extracted using the procedural pseudo-code reported in Algorithm-1. The Algorithm-1 process the DD in order to extract useful patterns at a certain usefulness value. The usefulness of the pattern present in the DD is validated by its TF-IDF threshold value. The higher value of TF-IDF of the news items is considered as useful items for the pattern as reported in line 3 of the Algorithm-1. The implemented framework also allows searching of certain pattern and/ or patterns from the extracted patterns.The extracted patterns can also be used for sentiment analysis, insight understanding of the text and determining trends of the news events. The sentiment polarities of the patterns are determined using SentiWordNet [39]. Dictionary and the computational scores are carried out using the similar method as mentioned in [16]. The proposed framework allows injecting any other method and/or technique for the knowledge discovery. The similar approach has been adopted in [37] for web navigational pattern extraction. On the contrary, the scientific research literature also reports several proposed methods to detect meaningful information from variety of data, such as text from video [40], clickstream data [37] relational database of criminal activities [38].

Block-4: Visual Representation: The extracted trends or patterns are presented in terms of effective visual graphs, which help in getting in-depth information about considered set of documents. The visual representation of information has potential to get in-depth understanding of considered dataset. For instance, study [38] proposed a framework to visually assess the trends of criminal activities, though the study applied such approach for structural data format.

The experimental results are discussed in the following subsection.

6.2 Results Discussion

This section reports the experimental results carried out at several events and topical models from news. The Fig. 4 reports word cloud and sentimental results for Malala Yousafzai (A young Pakistani girl, who is Nobile Prize winner) for a period of one-month news (24th March 2018 - 24th April, 2018).

The sentimental information about Malala over a month has been reported in Fig. 5. Malala has been in the news with a significant number of people in her favor, other against her and there are people who remain neutral about her stories. Likewise, the news regarding chief of opposition party PTI (Pakistan Tahreek-Insaf), a political party in Pakistan, remained in news in last two months (March and April, 2018) at different sections of the news such as sports, politics and religion as shown in Fig. 6.

The Fig. 7 reports the frequency of news in terms of words about Imran Khan. The visual representation of the extracted information from textual dataset (i.e. news and news articles) significantly portrait that how certain topics, people are given coverage in the news and news articles. This information helps in getting in-sight knowledge about media groups and their mindset, since media groups are the opinion makers of the society and they play key role in building societies and guiding them towards a national cause.

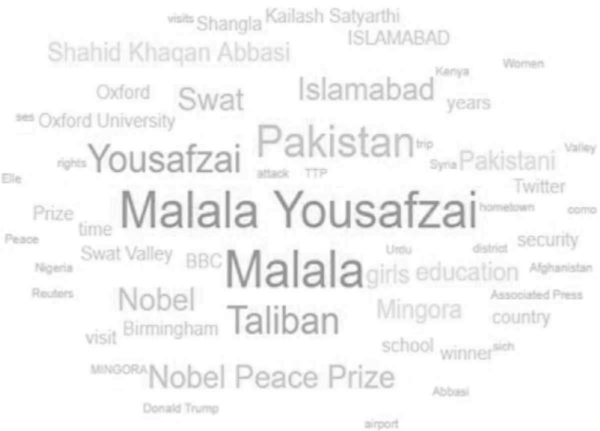

FIG. 4. WORD CLOUD FOR MALALA YOUSAFZAI

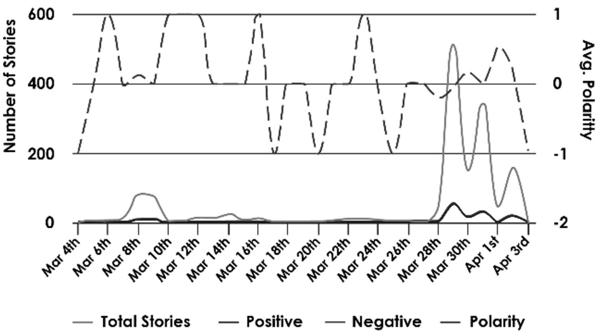

FIG. 5. SENTIMENTS OF PEOPLE FOR MALALA YOUSAFZAI

ALGORITHM-1 EXTRACT PATTERNS EXTRACT-PATTERNS (DD, MINTFIDF)

```
Require: Processed Document Database DD
Require: Minimum usefulness TFIDF threshold mintfidf
Ensure:  Extracted Patterns EP

for each news items $NT_k$ in DDdo
        Determine Categories and Conceptual terms - call AYLIEN-API($NT_k$)
                if TFIDF($NT_k$)≥mintfidf then Add $NT_k$ into EP
  endfor
return EP
```

The extracted information represented in effective manner helps in getting easy and potentially valuable knowledge about the patterns, trends and in-sights hidden in unstructured data formats. The above-mentioned results validate the effectiveness of the proposed framework that relies on basic text processing functions, thus, it requires less computational sources to bring shape to unstructured (i.e. textual) data.

7. CONCLUSIONS

TM with the technological growth has sought much attention among researcher community as well as business, political and media organizations. Undoubtedly, it is a challenging issue to uncover targeted and useful knowledge form natural language text. Many attempts have been done to cope with this challenging task.

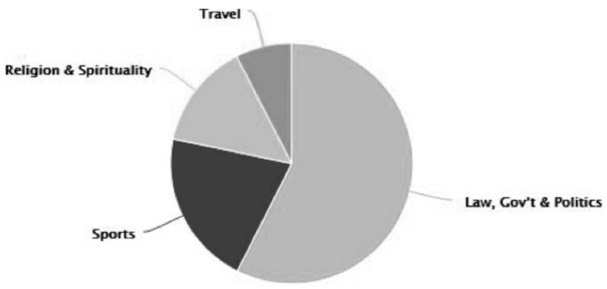

FIG. 6. THE NEWS CATEGORIES ABOUT IMRAN KHAN

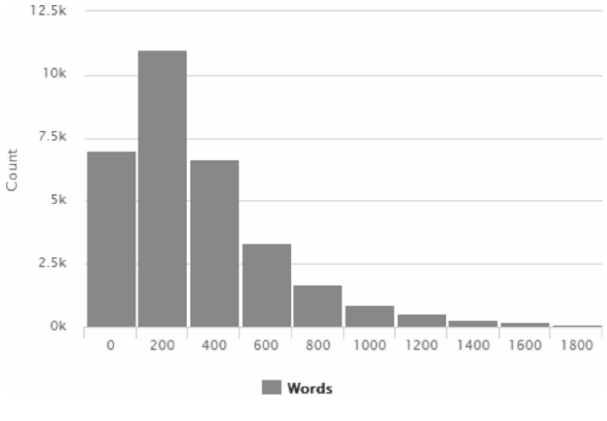

FIG. 7. THE WORD FREQUENCY ABOUT IMRAN KHAN

This study reported recent developments in the knowledge discovery from textual data. Determined inherent challenges in this area and proposed a framework that adopts basic NLP. The proposed framework extracts structures and trends from news articles; later, these trends are represented in the visual formats. The extraction of the trends have been done using well-established pattern mining technique (specifically AYLIEN Text Analysis API) and visual representation supports in getting better understanding of the insight knowledge hidden in the daily news and news article articles. The experiments have been performed over dataset of online daily news and news articles (English language).

The outcomes reveal the fact that how news events are covered in the electronic media. These trends highlight how daily news portraits the events, since media has been considered as opinion maker. Thus, trends assist in getting in-depth knowledge about media groups. The results may also help and assist government officials to regulate the media group based on certain facts discovered using proposed framework. The future work aims at extracting knowledge from audio and video streaming.

ACKNOWLEDGMENTS

The authors wish to thank continued support of the Institute of Information & Communication Technologies, Mehran University of Engineering & Technology, Jamshoro, Pakistan, for providing research facilities.

REFERENCES

[1] Schultz, J., "How Much Data is Created on the Internet Each Day?", 2018. (Online Blog - Last accessed on April 2018).

[2] Arron, L., Lyons, J., Akbari, A., Samantha, L.T., Angharad, M.W., Beata, F.-S., Owen, P., Mark, I.R., Ronan, A.L., David, V.F., and Rod, M.M., "Codifying Unstructured Data: A Natural Language Processing Approach to Extract Rich Data from Clinical Letters", International Journal of Population Data Science, Volume 333, pp. 1-38, 2017.

[3] Loechner, J., "90% of Todays Data Created in Two Years", 2018. (Online Blog - Last Accessed on April 2018).

[4] llangovan, R., "Big Data Analytics for Big Retail Success", 2018. (Online Blog - Last Accessed on April 2018).

[5] Tan, A-H., "Text Mining: The State-of-the-Art and the Challenges", Proceedings of Workshop on Knowledge Discoverery from Advanced Databases, pp. 65-70, 1999.

[6] Tonkin, E., and Tourte, G., "Working with Text: Tools, Techniques and Approaches for Text Mining", 1st Edition, Chandos Publishing, 2014.

[7] Vijayarani, S., and Janani, R., "Text Mining: Open Source Tokenization Tools - An Analysis", Advanced Computational Intelligence: An International Journal, Volume 3, pp. 37-47, 2016.

[8] Suresh, R., and Harshni, S.R., "Data Mining and Text Mining - ASurvey", International Conference on Computation of Power, Energy Information and Communication, pp. 412-420, March, 2017.

[9] Khan, K., Baharudin, B.B., Khan, A., and Malik, F., "Mining Opinion from Text Documents: A Survey", 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp. 217-222, June, 2009.

[10] Aparna, U.R., and Paul, S., "Feature Selection and Extraction in Data Mining", Online International Conference on Green Engineering and Technologies, pp. 1- 3, November, 2016.

[11] Zhang, Y., and Chen, M., and Liu, L., "A Review on Text Mining", 6th IEEE International Conference on Software Engineering and Service Science, pp. 681-685, September, 2015.

[12] Kaur, H., and Mangat, V., and Nidhi, "A Survey of Sentiment Analysis Techniques", International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 921-925, February, 2017.

[13] Talib, R., and Hanif, M.K., and Ayesha, S., and Fatima, F., "Text Mining: Techniques, Applications and Issues", International Journal of Advanced Computer Science and Applications, Volume 7, No. 11, pp. 414-418, 2016.

[14] Agbele, K.K., Ayetiran, E.F., Aruleba, K.D., and Ekong, D.O., "Algorithm for Information Retrieval Optimization", IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference, pp. 1-8, October, 2016.

[15] Liu, Y., and Shi, M., and Li, C., "Domain Ontology Concept Extraction Method Based on Text", IEEE/ACIS 15th International Conference on Computer and Information Science, pp. 1-5, June, 2016.

[16] Shaikh, A., Mahoto, N.A., Huda, N., Unar, M.A., "Discovering Twitter Sentiments Using Correlations Among Multiple Terms", Sindh University Research Journal (Science Series), Volume 48, No. 3, pp. 695-700, Jamshoro, Pakistan, 2016.

[17] Krishnan, A., and Sankar, A., Zhi, S., and Han, J., "Unsupervised Concept Categorization and Extraction from Scientific Document Titles", Proceedings of ACM Conference on Information and Knowledge Management, pp. 1339-1348, New York, USA, 2017.

[18] Aga, R.T., and Wartena, C., "Constructing Concept Clouds from Company Websites", Proceedings of 15th ACM International Conference on Knowledge Technologies and Data-driven Business, pp. 38:1-38:4, New York, USA, 2015.

[19] Xu, X.M., and Mutluand, Y.N., "Mining Concept Associations for Knowledge Discovery in Large Textual Databases", Proceedings of ACM Symposium on Applied Computing, pp. 549-550, New York, USA, 2005.

[20] Dinh, D., and Tamine, L., "Biomedical Concept Extraction Based on Combining the Content-Based and Word Order Similarities", Proceedings of ACM Symposium on Applied Computing, pp. 1159-1163, New York, USA, 2011.

[21] Chin, O.S., Kulathuramaiyer, N., and Yeo, A.W., "Automatic Discovery of Concepts from Text", IEEE/WIC/ACM Proceedings of International Main Conference on Web Intelligence, pp. 1046-1049, December, 2006.

[22] Abebe, S.L., and Tonella, P., "Natural Language Parsing of Program Element Names for Concept Extraction", IEEE 18th International Conference on Program Comprehension, pp. 156-159, June, 2010.

[23] Ajgalk, M., and Barla, M., and Bielikov, M., "From Ambiguous Words to Key-Concept Extraction", 24th International Workshop on Database and Expert Systems Applications, pp. 63-67, August, 2013.

[24] Szwed, P., "Concepts Extraction from Unstructured Polish Texts: A Rule Based Approach", Federated Conference on Computer Science and Information Systems, pp. 355-364, September, 2015.

[25] Yong-Bin K., Haghighi, P.D., and Burstein, F., "Cfinder: An Intelligent Key Concept Finder from Text for Ontology Development", Expert Systems with Applications, Volume 41, No. 9, pp. 4494-4504, 2014.

[26] Doan, P.T.H., Archint, N., and Archint, S., "Improving Key Concept Extraction using Word Association Measurement", 7th International Conference on Information Technology and Electrical Engineering, pp. 403-407, October, 2015.

[27] Khin, N.P.P., and Lynn, K.T., "Medical Concept Extraction: A Comparison of Statistical and Semantic Methods", 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 35-38, June, 2017.

[28] Wang, L., Fan, Z., Wang, X., and Yang, L., "Text Mining-Based Evaluation of the User Experience in Online Shopping for Clothing", 1stInternational Conference on Electronics Instrumentation Information Systems, pp. 1-4, June, 2017.

[29] Prameswari, P., Zulkarnain, Surjandari, I., and Laoh, E., "Mining Online Reviews in Indonesia's Priority Tourist Destinations Using Sentiment Analysis and Text Summarization Approach", IEEE 8th International Conference on Awareness Science and Technology, pp. 121-126, November, 2017.

[30] Jiang, M.,Pan, Z., and Li, N., "Multi-Label Text Categorization Using l21-Norm Minimization Extreme Learning Machine", Neuro-Computing, Volume 261, pp. 4-10, 2017, Advances in Extreme Learning Machines, 2015.

[31] Forman, G., and Kirshenbaum, E., "Extremely Fast Text Feature Extraction for Classification and Indexing", Proceedings of 17th ACM Conference on Information and Knowledge Management, pp. 1221-1230, New York, USA, 2008.

[32] Chintan, A., and Paauw, T., and Aly, R., and Lavric, M., "Identifying Child Abuse through Text Mining and Machine Learning", Expert Systems with Applications, Volume 88, pp. 402-418, 2017.

[33] Hashimi, H., and Hafez, A., and Hassan, M., "Selection Criteria for Text Mining Approaches", Computers in Human Behavior, Volume 51, pp. 729-733, 2015.

[34] Shihand, C.H., Yanand, B.C., Liuand, S.H., and Chen, B., "Investigating Siamese LSTM Networks for Text Categorization", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 641-646, December, 2017.

[35] Wang, F., Wu, W.,Li, Z., and Zhou, M., "Named Entity Disambiguation for Questions in Community Question Answering", Knowledge-Based Systems, Volume 126, pp. 68-77, 2017.

[36] Kaklauskas, A., Seniut, M., Amaratunga, D., Lill, I., Safonov, A., Vatin, N., Cerkauskas, J., Jackute, L., Kuzminske, A., and Peciure, L., "Text Analytics for Android Project", Procedia Economics and Finance, Volume 18, pp. 610-617, September, 2014.

[37] Mahoto, N.A., Memon, A., Memon, M.A., and Teevno, M.A., "Extraction of Web Navigation Patterns by Means of Sequential Pattern Mining", Sindh University Research Journal (Science Series), Volume 48, No. 1, pp. 201-207, Jamshoro, Pakistan, 2016.

[38] Shah, S., Khalique, V., Saddar, S., and Mahoto, N.A., "A Framework for Visual Representation of Crime Information", Indian Journal of Science and Technology, Volume 10, No. 40, pp. 1-8, 2017.

[39] Baccianella, S., Esuli, A., and Sebastiani, F., "Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", *Language Resources and Evaluation Conference*, Volume 10, pp. 2200-2204, 2010.

[40] Unar, S., Jalbani, A.H., Jawaid, M.M., Shaikh, M., and Chandio, A.A., "Artificial Urdu Text Detection and Localization from Individual Video Frames", Mehran University Research Journal of Engineering & Technology, Volume 37, No. 2, pp. 429-438, [ISSN: 2413-7219], Jamshoro, Pakistan, April, 2018.