Webology, Volume 16, Number 1, June, 2019

Investigation of Trends and Analysis of Hidden New Patterns in Prominent News Agencies of Iran Using Data Mining and Text Mining Algorithms

Babak Sohrabi

Professor, Faculty of Management, University of Tehran, Tehran, Iran.
ORCID: 0000-0001-6188-2607. E-mail: bsohrabi@ut.ac.ir

Iman Raeesi Vanani

Assistant Professor, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. ORCID: 0000-0001-8324-9896 E-mail: imanraeesi@atu.ac.ir

Meysam Namavar

MSc., Faculty of Management, University of Tehran, Tehran, Iran.
E-mail: meysamnamavar@ut.ac.ir

Received January 12, 2019; Accepted June 25, 2019

Abstract

These days, every business is trying to achieve its competitive goals. In other words, if one business uses the most modern analytical technologies optimally, it will certainly boost its success level at an exponential rate. The Islamic Republic News Agency (IRNA) is one such mass media industry that, according to its need, uses the benefits information technology (IT) provides it with. News flows, surveys, human resource management, warehouse management, software are key parts of the agency's news penetration criteria. There is little software that focuses on analytical solutions for running the business optimally. This paper, among available text mining methods, intends to present the most important keywords of news texts based on weighing of words and their correlations, news classification, news sentiments, news trends in order to provide insights and patterns for future scholars and practitioners in the field. This approach will help news agencies maintain their competitive edge as well as predict and react to the market of news search and provision in a timely context-based manner.

Keywords

News agency; Data mining; Text mining; Word correlation; Analytics

Introduction

News is “the communication of selected information on current events”, where the selection is guided by “newsworthiness” or “what interests the public”. News is also stories, from which the reader usually expects answers to the five Ws: who, what, when, where and why, to which a “how” is often added (Berendt, 2011). News and blogs consist of textual and (in some cases) pictorial contents, and, when Web-based, may contain additional contents in other format (e.g., video, audio) and hyperlinks. They are indexed by time and structured into smaller units: news media into articles, blogs into blog posts. In most news and blogs, textual content dominates. Therefore, text analysis is the most often applied form of knowledge discovery. This comprises tasks and methods from data/text mining, information retrieval, and related fields (Berendt, 2011).

Knowledge discovery in databases (Fayyad et al. 1996) is an iterative process of searching for valuable information, in large volumes of data, in a cooperative effort of humans and computers: humans select the data to be explored, define and analyze problems, set goals and interpret the results, while computers search through the data, looking for models and patterns that meet the human-defined goals (Fayyad et al., 1996). The central step in this process is data mining, the purpose of which is to automatically build classification models or find descriptive patterns in large data collections (Witten & Frank, 2005).

A variant of data mining is text mining where models and patterns are extracted from collections of text documents. In other words, text mining is relevant to linguistic research thanks to its ability to (i) process large amount of text, which is hard to do by hand, and (ii) automatically uncover non-obvious and unexpected patterns in language use, for example in newspaper discourse (Feldman & Sanger 2007).

Text mining techniques also include three major components of data mining i.e. classification, clustering, and association rule and sequence analysis. Among them, classification is a kind of form, which can be used to gather and describe important dataset (Han, Pei & Kamber, 2011). Using these techniques, the current paper tries to classify news, discover news with similar texts, analyze the sentiments of news and review the news trends. To further acquaint the readers, there is a brief description of the "news and its production", "text mining", "sentiment analysis", etc. Then the paper continues with "research proposal", "research method" and "discussion and conclusion".

News and news transmissions are of great importance in the contemporary era. In the past, there were only one news agency in some countries, but their numbers grew with dramatic increase in news agencies, newspapers and news sites worldwide. One of the main reasons was people’s appetite to know things and occurrences happening around them. As a matter of fact, news agencies are competing to attract people's attention for their own financial gains. In other words,

they are using knowledge, practices and new methods for gaining competitive advantage. Text mining is one of the subsets of the science of data mining that can be used by news agencies in their pursuit for set goals and policies.

Using text mining, the paper tries to reach the following goals:

- Since news items are categorized by manpower, time and effort, in a non-intelligent and non-systematic manner, and are stored in software called the news bulletin, this research aims to organize similar news by classifying algorithms and systematically store the relevant units for exploitation.
- Among the available text mining methods, weighing the words and their correlation are the most important and key to units responsible for analyzing the news.
- Repeating keywords over a specific time (one year) will help identify the news trend and provide for the relevant units.
- Using similarity of texts to find same or similar news in media agencies. In that case, the source can be quoted to avoid the copy right violation. Similarly, the search for texts can prevent plagiarism, authentication, displaying the place of similarity between arbitrary texts, retrieving information, and so on.
- Finding the sentiments of news.

Literature Review

Few years ago, just a handful of journalists knew that their core business - writing news - could be done so quickly by algorithms installed on newsroom computers. Part of the reason is that writing has long been taken as an "innately human" activity (Carlson, 2016, p. 228), an assumption not much challenged even when human-computer interaction becomes normal in the information age (Card, Moran & Newell, 1986).

The assumption, however, is seriously contested as algorithms are developed for writing news articles like earning reports, sports updates and breaking news. It is worth noting that today's algorithms can create and personalize an enormous number of articles much quicker, cheaper, and potentially with fewer errors than any journalist (Oremus, 2015). The algorithms can work round the clock in the newsroom, never needing a sick leave nor asking for a pay rise. More importantly, the overall quality of automated journalism keeps improving at a lightning speed, which may lead to big changes to the nature of journalism as a profession (see: Lewis, 2015).

What is News?

News is information about current events. As its name implies, "news" typically means the presentation of new information (Stephens, 1988; Smith, 1979). The newness of news gives it an uncertain quality, distinguishing it from the more careful investigations of history or other scholarly disciplines (Pettegree, 2014; Salmon, 1923; Park, 1940). While historians tend to view

events as causally related manifestations of underlying processes, news stories tend to describe events in isolation and exclude discussion of the relationships between them (Park,1940). In other words, news conspicuously describes the world in the present or immediate past, even when the most important aspects have occurred long in the past or are expected in the future. To make the news, an ongoing process must have some "peg", an event in time which anchors it to the present moment (Park, 1940; Schudson, 1986). Also, news often addresses aspects of reality which seem unusual, deviant, or out of context as a famous dictum suggests: "Dog Bites Man" is not news, but "Man Bites Dog" is (Park, 1940). Another aspect of the newness of news is that, as new technology enables media to disseminate news more quickly, 'slower' forms of communication may push it toward 'analysis' (Stephens, 1988).

Text Mining

The overall process of knowledge discovery in databases (KDD) is to identify valid, novel, potentially useful and ultimately understandable patterns (Fayyad, Piatetsky-Shapiro & Smyth, 1996). The term ‘knowledge discovery from textual databases’ (KDT) is a bit different from KDD and can be defined as discovering useful information and knowledge through the application of data mining techniques (Han, Pei & Kamber, 2011; Karanikas & Theodoulidis, 2002). However, it shares the common methods of collecting information as raw data and processing it. Moreover, in text mining techniques, a three step process of data collection, pre-processing and applications is required (Karanikas & Theodoulidis, 2002).

To perform text mining, it is necessary to modify texts to numeric so that data mining algorithms can be performed on large documents. As such, it requires the use of several techniques on texts whose scope varies from one word to the entire text database (Miner, Elder IV & Hill, 2012). In sum, the text mining methodology deals with removing noisy info inside unstructured data, extracting relevant ones as structured data and then applying algorithms. Meantime, it is necessary to use filtering techniques to reduce the space dimensionality since the space dimension is usually a central problem in statistical text clustering (Montoya, Villa, Muñoz, Arenas & Bastidas, 2015).

Sentiment Analysis

It has been widely observed that “what other people think” has always been an important piece of information for most of us during the decision-making process. Long ago, with the emergence of the World Wide Web, we usually asked our friends and colleagues to recommend a car mechanic or explain who they were planning to vote, called for job reference letters, or consulted to decide what washing machine to buy. With the Internet and the Web, we are no more in need for such consultations as they have now (among other things) made it possible to find out about opinions and experiences of people we never heard of. In other words, more and more people are making their opinions available online to strangers (Pang & Lee, 2008).

The sentiment analysis has emerged as a new field due to ever-increasing online reviews and data (Cambria, 2016). It aims is to analyze sentiments in written texts (Liu, 2015) but the number of possible applicants is much broader (Pang and Lee, 2008): business intelligence (analyze customer's reviews towards a product) (Mishne, et al., 2006), politics (predict election results, mining social opinions) (Wang, 2012; Bermingham, 2011; Ceron, 2014), tourism (Marrese-Taylor, et al., 2014; Valdivia et al., 2017), personality recognition (Majumder, et al., 2017) or social studies (evaluate the level of sexist messages on social networks or detect cyber bullying) (Jha & Mamidi, 2017; Chatzakou, et al., 2017; Xu, Bellmore & Zhu, 2012). While most studies approach it as a simple categorization problem, the sentiment analysis is actually a 'suitcase' research problem that requires tackling many subtasks of natural language processing (NLP), such as:

Sentiment Analysis Classification (SAC): This is the most popular task, whose aim is to develop models capable of detecting sentiment in texts. The sentiment is detected after collecting, reviewing and analyzing the text. It can be computed by the reviewer or with the specific anthropomorphic mannequins (SAMs). Then, features are selected to train the classification model, in which, text mining techniques are used to extract the most significant features (Valdivia et al., 2018).

Subjectivity Detection: This task is related to SAC since it aims to classify subjective and objective opinions. The purpose is to filter subjective sentences as they are more opinionated hence; can improve classification models.

Opinion Summarization: Also known as aspect-based summary or feature-based summary. It consists of developing techniques to sum up large amounts of reviews. The summarization should focus on entities or aspects and their sentiments and should be quantitative (Poria, et al., 2016; Hu & Liu, 2004).

Opinion Retrieval: This is a retrieval process, which requires documents to be retrieved and ranked according to their relevance. Sarcasm and Irony, for example, falls under this task which aims to detect opinions with sarcastic or ironic expressions. As in subjectivity detection, the target is to delete these opinions from the sentiment analysis process (Reyes, Rosso & Veale, 2013; Reyes, Rosso & Buscald, 2012). Since sentiment analysis is growing as a research branch, many new tasks have emerged over the past years e.g., temporal tagging (Zhong, Sun & Cambria, 2017), and word polarity disambiguation (Xia, Cambria, Hussain & Zhao).

Among the aforementioned tasks, SAC is important enough for sentiment analysis. Moreover, all the given models classify texts based on their sentiment, which can be identified as: label polarity (positive, neutral, negative), numerical rating (0, 2... 4 or 0, 1) or emotions (anger, disgust, fear, happiness, sadness, surprise) (Valdivia et al., 2018).

Related Works

The current paper is fully functional in the news field. What distinguishes it from others is that it explains all points and solutions encountered during the course of research.

Rip and Courtial (1984) were among the first to use the vocabulary analysis to describe the features of a scientific discipline. In a paper entitled "Drawing up a lexical biotechnology event", they reviewed articles published in a ten-year period in a biotechnology core journal. After encoding the terminology, they developed a coherent analysis of the vocabulary to provide a relationship between the texts, and then trends in the field of biotechnology were presented. These trends represent thematic relationships between the titles and their changes over time and according to the organization in which the research was conducted.

Callon et al. (1991), in their research entitled "The co-occurrence of words as a tool for describing a network of interactions….." used an analysis of vocabulary to understand the interaction between science and technology. In other words, they used this technique to link science and economics innovations. Their study was based on the content analysis of polymeric knowledge research information over a 15 year period, the outcome which showed that the academic research network was much bigger than the technological one, especially when the structure is changing faster than the applied research network. Meanwhile, they discovered that some of the topics were developed via scientific planning. Their findings also indicated that the simultaneous occurrence of vocabulary can be considered as a way to strengthen the interaction between academics and craftsmen. Chen (2004) categorized a variety of key points in the network and presented a method for representing the network's time course, but in his research he neither paid attention to the evolution of the network nor provided a model for the recognition of new trends.

Kim and Lee (2008) investigated 423 articles published between 2001 and 2004, and examined 43 clusters of documents in the software service system (SSS-software). Then a grid network built these 43 clusters in seven subject areas: digital libraries and digital documentation technology, online resources and hedge funds, archives and archivists, political and legal issues, technical issues and electronic records, management information and records, and email and professional information. Finally, these seven issues are integrated into three sections i.e. digital libraries, archives and research. This study shows a dynamic variation in research topics from a purely traditional subject area to complex ones from 2001 to 2004. The results also suggest that researches in the field of legal science are capable of growing and will continue to develop. Cozzens et al. (2010) presented technology growth indicators based on different perspectives and examined quantitative methods for identifying new areas of technology, but did not provide a model for recognizing these domains. Yang et al. (2011) conducted a research entitled "Thematic Analysis of Hospital Care Research Using Vocabulary Coincidence", with the aim of mapping

the structure of the hospital care system using the occasional vocabulary method. They used nerve networks in text mining to map the area of interest. They came to the conclusion that health care and services as two domains vital in studies related to improved patient care. Guo, Weingart and Börner (2011) described the emerging mechanisms of new phenomena and provided a model for diagnosis, but they focused on networks separately in their research and did not provide a comprehensive model for prediction.

Wührer et al. (2011) used a combination of network analysis and identified new marketing domains in Turkey using scientometrics. Their research basically focused on the current state of Turkey and lacked a model for predicting the future. Tu et al. (2012) presented a novel model but it not only takes a long time to diagnose but cannot be used for all new trends. Akritidis et al. (2012) presented a model for discovering appropriate fields for young researchers but it focused solely on individual characteristics of the trainees and did not include comprehensive criteria for evaluating the results. Alexander et al. (2012) presented a new and emerging concept and attempted to scale this to the system and subsystems, but they too failed to provide a useful index for high-volume information analysis.

Pabalkar (2012) conducted a research on classifying news and techniques. Cheney (2013) explored the opportunity for scholars to publish newspaper articles and news content free of charge. Small et al. (2014) presented a method to explain new domains in large volumes of information that their research focused on. Boyack et al. (2014) presented a model for identifying mechanisms for changing clusters. Their research problem focused on clustering. Zhu and Yu (2014) presented a micro-blogging model to determine hot topics that focused on word extension and lacked attention to other content structures. Wang et al. (2014), focusing on the network, presented a model vocabulary that analyzed the research domains based on dynamic vocabulary.

Khadjeh Nassirtoussi et al. (2014) conducted a research to predict the Forex market using news-related text mining headline, especially for two reasons. First, more emphasis was placed on the use of text mining techniques, especially in order to complete the problems encountered in previous research. Second, the research had focused on a certain range of foreign exchange markets that had not been previously explored. The outcome demonstrated the relationship between a particular type of market and textual data from the news and the challenges it poses. Kim et al. (2016) used online electronic articles to investigate and find relationships between words. The model used in this study was a fuzzy cognitive map. Al-Anazi et al. (2016) manually compiled articles from the data collection at the Royal University of Saudi Arabia, and then, using the K-Mean algorithms for clustering of articles to find the best one.

Table 1. Related works

Researchers	Year	Issue	Goal
Rip & Courtial	1984	Co-word maps of biotechnology: An example of cognitive scientometrics. Scientometrics	Providing trends in the field of biotechnology and finding links to texts
Callon et al.	1991	Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry	Showing an analysis of both the occurrence of vocabulary can be considered as a way to consolidate the factor between academics and craftsmen
Chen	2004	Searching for intellectual turning points: Progressive knowledge domain visualization	Providing a method for displaying network time
Kim et al.	2008	Exploring the emerging intellectual structure of archival studies using text mining	Grouping topics in seven subject areas
Cozzens et al.	2010	Emerging technologies: quantitative identification and measurement	Investigating quantitative methods for the detection of new technology domains
Yang et al.	2011	The topic analysis of hospice care research using co-word analysis and GHSOM. Intelligent Computing and Information Science	Outlining the scope of hospital care
Guo, Weingart & Börner	2011	Mixed-indicators model for identifying emerging research areas	Providing a model for identifying new trends
Tu & Seng	2012	Indices of novelty for emerging topic detection. Information processing & management	Finding new trends
Akritidis et al.	2012	Identifying attractive research fields for new scientists. Scientometrics	Discovering new and appropriate research areas
Alexander et al.	2012	Emergence as a conceptual framework for understanding scientific and technological progress. Paper presented at the Technology Management for Emerging Technologies	Leveling the concept of visibility into the system and subsystems
Sarika Y.	2012	Classifying news using the web of text mining	Classifying news
Cheney	2013	Text mining newspapers and content News: New research trends and methodologies	Discovering knowledge from news texts
Small et al.	2014	Identifying emerging topics in science and technology. Research policy	Explaining new areas in large volumes of information
Boyack et al.	2014	Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science	Reviewing of new research areas
Zhu & Yu	2014	A pre-recognition model for hot topic discovery based on microblogging data. The Scientific World Journal	Finding new and hot topics
Wang et al.	2014	Analyzing evolution of research topics with viewer: a new method based on dynamic co-word networks	Providing a model for analyzing research fields based on dynamic vocabulary
Khadjeh Nassirtoussi et al.	2014	Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment	Predicting the relationship between foreign market prices and textual data from headline news
Kim et al.	2016	Futuristic data-driven scenario building: incorporating text mining and fuzzy association rule mining into fuzzy cognitive map	Finding relationships between words
Al-Anazi et al.	2016	Finding similar documents using different clustering technique	Finding the best algorithms for clustering and compare them with each other

Materials and Methods

The method applied in the course of study is considered as one of the important parts of any research. It gives an insight as how data were obtained and interpreted since the applied method affects the findings. In other words, methodology is also crucial because an unreliable one produces unreliable results. CRISP-DM is the most popular data mining methodology, which breaks the data mining process into six major phases:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

This methodology can be used for text mining researches, where text not data context is taken into account. With this, the current study was conducted in four steps:

- Extracting news items from news agencies' websites and integrating as well as storing them in a database (IRNA, Fars, EghtesadOnline, and Tasnim)
- Preparation and cleansing of news texts
- Implementation of text mining algorithms such as weighing words and classification in RapidMiner
- Finding new trends in the development and implementation of information systems.

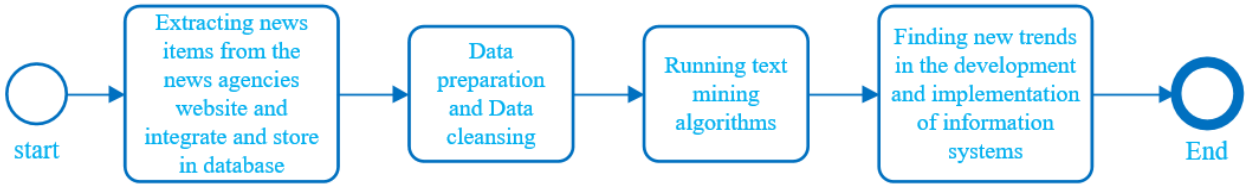

Figure 1. Steps of conducting research

The first step: Extracting news items from the news agencies' websites and integrating and storing in a database (IRNA, Fars, EghtesadOnline, and Tasnim)

For the text mining, we must first provide some news for which, the easiest way possible is to have contact with news agencies or request access to their databases. But a news agency is unlike to allow access to its database. Since some sorts of news were needed to go ahead with processing and analysis, we could use *web scraping* to extract and collect the news. In short, web scraping is a process that automatically collects data from websites. As a matter of fact, we began collecting data through web-based software fully relevant us. These data are categorized as follows:

- URL link of news
- Date and time of the publication of news

- News code
- News headline
- News lead
- News body
- News categorization

The data needed to be stored following their extraction. Due to access speed and increased flexibility in processing and analysis, news items were stored in Microsoft SQL Server databases. From the beginning, the program was designed to avoid processing and preparation for the news. All aforementioned news items were stored in the database, with tables and independent columns showing for each news agency.

The second step: Preparation and cleansing of news texts

Although each new item was stored independently but they need to be preprocessed leading to the removal of non-news phrases from the texts.

In this step, the words or sentences in the text that are not related to the news itself are deleted. For example:

> *'We had had good talks with Maersk Oil, which were close to being finalized, but Total bought the company', said Bijan Zangeneh.*
> *He said that the oil field is now producing 20,000 barrels a day; Iran has made requests for its development but has not received any serious offers, yet.*
> *9417**1396*
> *Follow us on Twitter @IrnaEnglish*

The above news at the end contains a code and a Twitter URL for the IRNA English. This code corresponds to a journalist whose news has been published. Obviously, this information is not self-reported and does not help in analyzing the news, so they must be removed before processing the texts of all the news.

Tokenizing: The process of breaking a stream of textual content up into words, terms, symbols, or some other meaningful elements is called tokens.

Stemming: This is a process for reducing modulated words to their stem, root or base.

Spelling integration: In this step, spellings that are different but still have the same meaning, should be identical to get better results such as, organization.

N-grams: These are extensively used in text mining and natural language processing tasks. They are basically a set of words within a given window and when computing the n-grams, you typically move one word forward (although you can move X words forward in more advanced scenarios). For example, for the sentence "*The cow jumps over the moon*".

If N=2 (known as bigrams), then the n-grams would be:

- *the cow*
- *cow jumps*
- *jumps over*
- *over the*
- *the moon*

Filtering: This step involves filtering stop words from within the text, and also filtering some tokens from within the text.

The third step: Implementation of text mining algorithms such as weighing words, classification and ... in RapidMiner

At this stage, weighing all words using the TF-IDF method, which is the best and, at the same time, the most sophisticated algorithm. In fact, TF-IDF is a weight that is used in data retrieval and text mining, which is a static scale evaluating the importance of a word in a document from a bundle of documents. This is proportional to the number of repeated words in the texts, but is balanced by the amount of word occurrence in the whole text parsing. It should be noted that in this research all of the text mining processes were performed using "RapidMiner" software. All processing is done via text classification algorithms, similarity detection, sentiment analysis, etc.

The fourth step: Finding new trends in the development and implementation of information systems

Finally, the output of the created models should be used in the organization's existing information systems to achieve the intended purpose.

Results

The most important part of any research is the analysis of data collected in order to answer the research questions. In this section, using text-mining techniques, we try to classify the news, find news trends over a certain period (Year 2017), similar news and sentiments analysis of news texts. The algorithms used in this study are the most promising and efficient in their domains. In a field like classification of texts that have several algorithms, even four of the most popular classification algorithms were used and evaluated to be used most effectively for this research purposes.

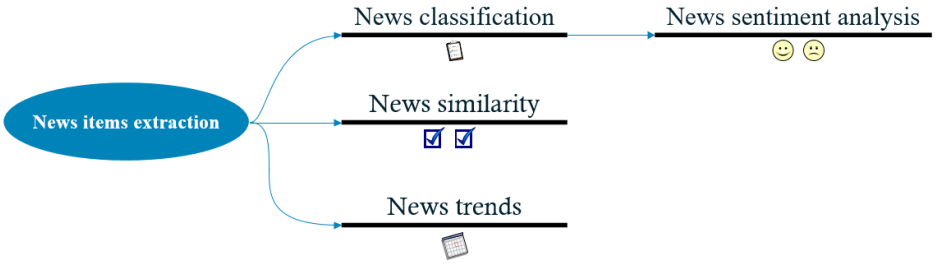

Figure 2. An overview of research findings

As shown in Figure 2, text and data mining algorithms were used after extracting the news items (News link, News code, News headline, News lead, News body). The classification will categorize the news based on needs of the analysis team of news agencies. Thereafter, the categorized ones will be classified based on positive or negative sentiments; hence, it increases the speed of producing news analytics.

Plagiarism is a "wrongful appropriation", "stealing" and publication" of another author's "language, thoughts, ideas, or expressions" and the representation of them as one's own original work. Since journalism relies on the public trust, a reporter's failure to honestly acknowledge his sources undercuts the integrity of a newspaper or television and undermines its credibility. Journalists accused of plagiarism are often suspended from their reporting tasks while the charges are being investigated by the news organization. News similarity will be used for identifying two or more similar news all over the agencies websites. It will help us identify news plagiarism. News trends will also help to discover strength and weakness of national issues based on their reporting over a specific period of time that are produced by the news agencies.

1. News Items Extraction

As explained earlier, news should be automatically collected from all news agencies' websites and stored in databases. So a web application was developed using the ASP.NET framework. The task of this program is to crawl on news stories (IRNA, Fars, Tasnim, and Eghtesad Online) and read all the contents of HTML pages and save items to the database.

The items on every news site are as follows:

- **Date and time of news release**
- **News code:** A unique code for any news
- **News headline:** The most important news message is summarized and compressed to encourage readers to go for the lead and the body.
- **News lead:** Leads are a summary of the most important news section. In fact, a lead is nothing but our first few statements that guide the audience to the news. A good lead contains 10 to 15 words.
- **News body:** Full description of the news.

It will be explained below which news items is better for achieving the goals, such as classifying

news stories, reviewing news trends, analyzing news sentiments and discovering news with similar texts.

Table 2 briefly presents the number of news collected from news agencies (IRNA, ISNA, FARS, and Eghtesad Online).

Table 2. Statistics of News Agency of Research

Name	Politics	Economic	Sport	Art & Culture	Total
IRNA	5100	1652	519	1364	8642
Fars	10206	499	10	245	14881
Tasnim	852	437	-	454	1816
Eghtesad Online	110	8180	-	-	8290
+ISNA	-	-	-	-	-

Since, the news items will not remain for more than two days on the ISNA website, it is not possible to extract them. Some news agencies are specialized in a particular area. For instance, the Eghtesad Online specializes in economics and does not publish news in other areas such as sports and culture. Table 3 shows the number of items published by news agencies given their respective areas (political, economic), with a significant share in the news production.

2. News Classification

News from agencies and news websites are classified into different categories that are not necessarily identical with other agencies. In this paper, they are classified into the following four categories:

- Political news
- Economic news
- Sports news
- Artistic and cultural news

Every news item is provided by a correspondent working in social, political, economic, sports, filed, and is referred to a consultant or editor for review and approval. Once verified, it will be uploaded by the news agency. But the news itself will not be worth it with the increasing number of agencies or websites and even social networks because it is easily accessible to everyone. Of course, one of the goals of news agencies everywhere in the world is to make people aware about issues. But if one can get the knowledge of this huge volume of news being produced and published, the value of news will be multiplied. One of the applications that can be used for text mining in the news domain is to categorize news feeds and other sites for analysis and research units.

Nowadays, analysts separate and study the news and articles according to their specialized fields

(non-systematic). By categorizing news stories, feeds can be automatically provided to them daily, and they can dramatically increase the speed of separation and downloading and analyzing the items.

Since, news algorithms classification requires model training, algorithms must first be taught by a series of news stories, with the aim to create a model. To train a model that can categorize the news, there are several algorithms in which, the following are presented in this paper for teaching and testing the model.

Text mining and models can be used to categorize news (economy, political, sports, culture and art) depending on their need for a headline, feed or text. Classification algorithms, first, creates a learning model and then tests its validity. Since the headline goes to the body, the number of words goes up as well. Therefore it can be said, more the words, the better the model is.

In the following two tables, one is created with training and test data in X-Validation by RapidMiner software. The second table relates to testing algorithm's performance on other news sources, such as those from Tasnim and Fars, to check the reliability of the model. It should be noted that for training the models, 2,600 news items from IRNA were used. In fact, for each of the four news (political, economic, sports, art-culture), 650 items are intended to train the model.

The following table relates to the model training, using SVM, K-NN, Decision Tree, and Naive Baysian algorithms. This training was conducted using news from IRNA.

Table 3. Performance of algorithms using IRNA news

Algorithm	Accuracy
SVM	87.31% +/- 1.66%
K-NN	84.92% +/- 0.73%
Decision Tree	48.27% +/- 0.40%
Naive Baysian	25.00% +/- 0.00%

After training, the model should be evaluated with the news from other agencies in order to ensure its correctness and performance. The following table describes the performance of each of the four algorithms, using reports from Tasnim and Fars news.

Table 4. Test performance algorithms using Tasnim and Fars news

Algorithm / News Agency	SVM	K-NN	DecisionTree	NaiveBaysian
Tasnim (Top 20% news record)	81.05%	71.80%	27.67%	2.05%
Fars (Top 20% news record)	88.02%	79.52%	-	-

Tip 1: There is a need to balance the number of news in different categories. Otherwise, modeling will have the most news. For example, the number of political news has the most in all agencies. Now, if we allocate all for teaching, the model's accuracy for classifying other categories is greatly reduced because it overcomes the others. Therefore, to train the model from the category that had the smallest number of news at IRNA news agency, the sports field was used.

Tip 2: The neural network algorithm was also evaluated for categorization so that even after a while, the algorithm was not able to provide a model; the modeling process was inevitably canceled.

As shown in Tables 3 and 4, the "Support Vector Machine" algorithm, after training and testing all other algorithms, achieved a much higher output. After processing the news texts in Rapid Miner software, the output of the text processing was provided by SVM algorithm. A model with 87 percent accuracy was trained by this algorithm. The model was then evaluated for measuring its performance with the news of other agencies such as Tsunim and Fars, which resulted in 81.05 percent and 88.02 percent, respectively. This is an extraordinary performance.

The second algorithm used to classify the news is the K-Nearest Neighbor algorithm. This algorithm is strongly dependent on K parameter. To determine and calculate the best value K, there is no scientific method except to increase and decrease it to reach the desired result. For this reason, K was set to 1 to start up to an optimal value of 65. More or less, it was classified by the news with lower accuracy.

The accuracy of the K-Nearest Neighbore algorithm was 84.92 percent, but the point here is that the accuracy of the model test, which was done using the news reports of Tasnim and Fars, is far from the accuracy. The accuracy of the model test is 71.80 percent and 79.52 percent respectively for Tasnim and Fars news. Meanwhile, the news was also used to test the model produced by SVM and K-NN algorithms.

The timing of the decision-tree algorithm is time-consuming relative to the two previous algorithms, and consumes more resources from the system based on the logic and structure of the decision-tree algorithm. The expected tree decision was not achieved in the category of news classification so that the training accuracy of the model was 48.27 percent. The accuracy of the model test was also 27.67 percent with the Tasnim news agency's news, which is not at all an appropriate. Naive Baysian algorithm, like the Decision-Tree algorithm, does not have the proper function for our research purpose. The accuracy of training and testing the model is 25 percent and 2.05 percent, respectively.

3. News Trends

The trend of news is another application of text mining. It can also be useful for news analysis

and research units. The trend in the headlines of IRNA, Fars, Tasnim, Eghtesad Online has been reviewed in the first half, the second half and the entire 2017. But for the sake of summary, only 10 full-repeat words in 2017 are described for each news agency. A noteworthy and remarkable thing is the words "Iran", "Oil" and "Export" that are on the top of the table and indirectly depend on the country's macro economy for oil. In the search for news trends, headlines were processed and analyzed and their modeling was used. Since the headline is short news where keywords are used, finding the number of repetitions of these keywords, we will get the news trend. As said, the results of the news trend were based on three repeated words "Iran, Oil, Export." This highlights Iran's news stance as confirming its reliance on oil exports. Now, if this model is implemented on the items of the news agencies of other countries, their trends and attitudes are also clear to the research groups.

In general, there is not much difference in the use of news headlines and news leads in modeling because the headline lead has more descriptions on a scale of up to 40 words. Of course, it was said that the best lead should contain 10 to 15 words. The results of the news trends are presented in separate tables.

Table 5. Fars news trends

Word	Attribute Name	Total Occurences	Document Occurences
Iran	Iran	421.0	421.0
Oil	Oil	86.0	86.0
Export	Export	63.0	62.0
Trade	Trade	53.0	53.0
Minist	Minist	52.0	52.0
Offici	Offici	51.0	51.0
Ink	Ink	44.0	44.0
Cooper	Cooper	43.0	43.0
Gas	Gas	39.0	39.0
South	South	35.0	35.0

Table 6. Tasnim news trends

Word	Attribute Name	Total Occurences	Document Occurences
Iran	Iran	352.0	351.0
Oil	Oil	94.0	92.0
Gas	Gas	60.0	60.0
Offici	Offici	60.0	60.0
Iranian	Iranian	59.0	59.0
Deal	Deal	53.0	53.0
Minist	Minist	52.0	52.0
Export	Export	45.0	45.0
Report	Report	40.0	40.0
Sign	Sign	34.0	34.0

Table 7. IRNA news trends

Word	Attribute Name	Total Occurences	Document Occurences
Iran	Iran	1100.0	1089.0
Export	Export	178.0	178.0
Oil	Oil	171.0	159.0
Cooper	Cooper	150.0	150.0
Offici	Offici	150.0	150.0
Trade	Trade	128.0	125.0
Minist	Minist	122.0	122.0
Tie	Tie	119.0	119.0
Iranian	Iranian	112.0	112.0
Econom	Econom	98.0	97.0

Table 8. Eghtesad Online news trend

Word	Attribute Name	Total Occurences	Document Occurences
Iran	Iran	2740.0	2727.0
Oil	Oil	665.0	653.0
Export	Export	632.0	632.0
Bank	Bank	497.0	486.0
Trade	Trade	463.0	462.0
Iranian	Iranian	427.0	427.0
Market	Market	353.0	352.0
Tehran	Tehran	296.0	295.0
Import	Import	287.0	287.0
Rise	Rise	274.0	274.0

4. News Similarity

We use texts to find the same or similar news in news agencies. This topic can be found in the news published by a news agency and re-issued by another agency with or without the source. Thus, the copy right of the news can be proven to the source of the news agency. Similarly, the search for texts can prevent plagiarism, authentication, displaying place of similarity between arbitrary texts, retrieving information, and so on.

In the news similarity, texts were used since we were looking for the most similar ones, and the headline or lead cannot respond to such request because of fewer words, and finding similarity between several words is not very logical. By using the time of news release, it is also possible to find out which ones were sooner on their outputs, and possibly if the copying was done, the originality of the news relates to which agency.

In order to investigate this issue, 25 news stories from IRNA and 25 from Tasnim were compared to find similarity. The table below is one of 50 news items selected for similarity. In the table below, the similarity value obtained is equal to 67.5 percent.

Table 9. Table of characteristics of similarity of two news

First News Row	Second News Row	Similarity
https://www.tasnimnews.com/en/news/2018/02/07/1650406	http://www.irna.ir/en/news/82823017/	0.675340212

Figures 2 and 3 show relevant news on IRNA and Tasnim websites. The effectiveness of this idea in both cases is that even the pictures of the two have the same images and are identical.

Figure 3. Similar news in IRNA

Figure 4. Similar news in Tasnim

After processing the news texts in RapidMiner software, the words are scored based on the TF-IDF algorithm. This score is numerically the input of the similarity algorithm. Here, the similarity measurement type is numerical measures and the numerical measure is cosine similarity.

Note: In similarity analysis, the sources involved are strongly linked to the fact that news texts are cross-linked. For, 25 news stories from two agencies were dealt with, and if all the news were to be examined, we would surely find it above the mentioned similarities and percentage.

5. News Sentiments Analysis

Newsgroups focus on analyzing negative or positive news in their editorials. This is done manually and time consuming. We plan to identify the post-auto-news categories, negative or positive news, in order to increase the speed and accuracy of analytical news production.

Tip 1: Given the indescribable flexibility of natural language, the sentiment analysis with existing algorithms remains at the same level as that of machine diagnosis, and it cannot be said with certainty that a text is completely positive or negative. This is also because metaphors and caricatures are not recognizable to the machine. But as we can filter the news to some extent and reach the right person at the right time, we somewhat reached our goal.

Tip 2: Throughout the research, we found that sentiment analysis algorithms to detect the polarity of some news categories, such as economic and sports are severely inadequate because the texts of these newsgroups contain numbers (economic news) or figures (sports news).

The Aylien plug-in has been used in RapidMiner software to analyze the news sentiments. This plug-in uses natural language processing to distinguish and analyze the sentiments. The plug-in provides up to 1000 records per day for use by researchers, but for a larger number, one will need to purchase a license from the Eileen website. In the sentiment analysis, news bodies were also used for modeling, because the headline and the lead lack enough characters to detect the positivity or negativity of a sentence. In this research, 200 news items in the fields of "politics" and "culture and art" have been evaluated. In the table below, four positive and negative polarities are listed for a sample survey.

Table 10. News sentiment analysis

Headline	Link	News category	Polarity	Polarity confidence
President Rouhani: Iranians not needing sympathy of ill-wishers	www.irna.ir/en/News/82781920	Politics	Negative	1
8 minor quakes felt in SW Iran	www.irna.ir/en/News/82781031	Arts & Culture	Negative	0.997290219
President Rouhani visits Christian martyrs' families	www.irna.ir/en/News/82774754	Politics	Positive	0.999532407
Iranian kids win Mental Arithmetic Champs	www.irna.ir/en/News/82780414	Arts & Culture	Positive	0.901763872

The above table shows the output of the Aylien plug-in in RapidMiner software, and contains news headline, link to access the news, news categorization, the polarity, and the degree of confidence in the polarity of the news. In other words, the degree of polarity is 0 to 1, meaning the closer one is, the more positive news is confidently detected.

Conclusion

With the advent of knowledge and technology, the organizations and companies that benefit most are on the forefront. Articles are also being developed to achieve this goal in a way that can contribute to the growth of organizations and companies. Today, there is a growing competition in the news arena worldwide, with news sites becoming a substitute for print media especially on cultural development front as well as raising general awareness and pursuing people's rights. Also the presence and activities news agencies are being overshadowed with the emergence of relatively small news websites and social media networks. As said before, this approach will help the news agencies maintain their competitive edge over their competitors. Some important news from around the world provide feeds for news analysis units, but the text mining news can discover important issues in the news and provide them with analysis units, especially when

there is no important news for analysis. Text mining is one of the subsets of data mining that can be used to meet the objectives and policies of the agency. The purpose of this study was to evaluate the new trends in the development and implementation of information systems, using text mining techniques. The most important results we tried to achieve were:

- Exploiting the science of the day to achieve organizational goals more efficiently
- Maintaining the competitive advantage of an enterprise with respect to modern technologies
- Extending previous research in the field of text mining to achieve more functions of this knowledge
- Assist in the provision of enterprise-based software based on data mining and text mining for the independence of search engines

There are some points that future researchers should consider:

First, this research is done by RapidMiner, a production tool that works in the data mining domain. This is not programming language software hence, cannot produce a program or join an organization's information systems. It is suggested, however, that Python which is one of the best programming languages can be used to carry out researches in this field.

Second, text mining is a complicated process and is involved with system resources. For running this process in high quality, researchers need to use a proper hardware system.

The application of this study should enable news organizations to quickly and more accurately analyze the news, and convert the value-added items from "informing" to "news with comprehensive and applied analyzes and reports". In other words, the users can get a detailed look at the analysis and commentary of the news in each area and use them in their own field of work.

References

Akritidis, L., Katsaros, D., & Bozanis, P. (2012). Identifying attractive research fields for new scientists. *Scientometrics*, 91(3), 869-894.

Al-Anazi, S., AlMahmoud, H., & Al-Turaiki, I. (2016). Finding similar documents using different clustering techniques. *Procedia Computer Science*, 82, 28-34.

Alexander, J., Chase, J., Newman, N., Porter, A., & Roessner, J. D. (2012). Emergence as a conceptual framework for understanding scientific and technological progress. In *2012 Proceedings of PICMET'12: Technology Management for Emerging Technologies* (pp. 1286-1292). IEEE.

Banfield, A. (2014). *Unspeakable sentences: Narration and representation in the language of fiction*. Melbourne: Routledge.

Berendt B. (2011) Text mining for news and blogs analysis. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning. Springer*, Boston, MA, pp. 968-972.

Bermingham, A., & Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pp. 2-10.

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management*, *32*, 147-159.

Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1), 155-205.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31 (2) 102-107.

Card, S. K., Moran, T. P., & Newell, A. (1986). *The psychology of human-computer interaction*. Ahawah, New Jersey: Lawrence Erlbaum Associates.

Cardie, C. J. W. (2003). Combining low-level and summary representations of opinions for multi-perspective question answering. *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, pp. 20-27.

Carlson, M. (2016). Automated journalism: A post-human future for digital news? In B. Franklin, & S. A. Eldridge (Eds.), *The Rutledge Companion to Digital Journalism Studies*. London: Routledge (pp. 226-234).

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, 16(2), 340-358.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, June). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 13-22), June, 2017. ACM.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences*, 101 (Suppl 1), 5303-5310.

Cheney, D. (2013). Text mining newspapers and news content: New trends and research methodologies. *IFLA WLIC 2013*, Singapore. Retrieved January 15, 2019, from http://library.ifla.org/233/1/153-cheney-en.pdf

Cozzens, S., Gatchair, S., Kang, J., Kim, K.-S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: quantitative identification and measurement. *Technology Analysis & Strategic Management*, 22(3), 361-376.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful Knowledge from volumes of data. *Communication of the ACM*, 39 (11), 27-34.

Feldman, R., and J. Sanger (2007). *The text mining handbook. Advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Guo, H., Weingart, S., & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, 89(1), 421-435.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. San Francisco: Elsevier.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. *Proceedings of the SecondWorkshop on NLP and Computational Social Science*, 7–16.

Khadjeh Nassirtoussi, A., Aghabozorg, S., Wah, T.Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-Layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1), 306-324.

Kim, H., & Lee, J. Y. (2008). Exploring the emerging intellectual structure of archival studies using text mining: 2001-2004. *Journal of Information Science*, 34(3), 356-369.

Kim, J., Han, M., Lee, Y., & Park, Y. (2016). Futuristic data-driven scenario building: incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. *Expert Systems with Applications*, 57, 311-323.

Lewis, S. C. (2015). Journalism in an era of big data: Cases, concepts, and critiques. *Digital Journalism*, 3(3), 321-330.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2), 74-79.

Marrese-Taylor, E., Vel'asquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews, Expert Systems with Applications 41 (17) 7764–7775.

Miner, G., Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications: Academic Press.

Montoya, O. L. Q., Villa, L. F., Muñoz, S., Arenas, A. C. R., & Bastidas, M. (2015). Information retrieval on documents methodology based on entropy filtering methodologies. International Journal of Business Intelligence and Data Mining, 10(3), 280-296.

Negrine, R. (2003). *Politics, and the mass media in Britain*. London: Rutledge, pp.137-162

Nikhil, R., Tikoo, N., Kurle, S., Pisupati, H. S., & Prasad, G. (2015). A survey on text mining and sentiment analysis for unstructured web data. Paper presented at the Journal of Emerging Technologies and Innovative Research

Oremus, W. (2015). Why robot? Automated journalism is no longer science fiction. It's time to change what we call it. *Slate*, February 5, 2015. Retrieved January 15, 2019, from http://www.slate.com/articles/technology/future_tense/2015/02/automated_insights_ap_earnings_reports_robot_journalists_a_misnomer.html

Pabalkar, S. Y. (2012). Web text mining for news by classification. *International Journal of Advanced Research in Computer and Communication Engineering*, 1(6), 387-391.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

Park, R. E. (1940). News as a Form of Knowledge: A Chapter in the Sociology of Knowledge. *American Journal of Sociology*, 45(5), 669-686.

Pettegree, A. (2014). *The invention of news: How the world came to know about itself*. New Haven, Conn.: Yale University Press.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108, 42-49.

Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: a business oriented approach. Expert Systems with Applications, 39 (5), pp. 4729 – 4739

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, *74*, 1-12.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1), 239-268.

Rip, A., & Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6), 381-400.

Salmon, L. M. (1923). *The newspaper and the historian*. Oxford: Oxford University Press.

Schudson, M. (1986). When: Deadlines, datelines, and history. In R. K. Manoff Si M. S. Schudson (Eds.), *Reading the News* (pp. 79-108). New York: Pantheon.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. Research Policy, 43(8), 1450-1467.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, Volume 1, pp. 1555-1565.

Tu, Y.-N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management*, 48(2), 303-325.

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment Analysis in TripAdvisor. *IEEE Intelligent Systems*, 32(4), 72-77.

Valdivia, A., Luzón, M. V., Cambria, E., & Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion*, *44*, 126-135.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. *Proceedings of the ACL 2012 system demonstrations* (pp. 115-120). Association for Computational Linguistics.

Wang, X., Cheng, Q., & Lu, W. (2014). Analyzing evolution of research topics with NEViewer: A new method based on dynamic co-word networks. *Scientometrics*, 101(2), 1253-1271.

Witten, I.H., & Frank, E. (2005). *Data mining practical machine learning tools and techniques*. 2nd Edition.. San Francisco: Elsevier.

Wührer, G. A., Bilgin, F. Z., & Karaosmanoğlu, E. (2011). *The development and transfer of scholarly marketing thought in Turkey: A scientometric analysis of master and PhD theses 1980–2008 in an emerging market country*. Bilgi Ekonomisi ve Yönetimi Dergisi, 6(1).

Xia, Y., Cambria, E., Hussain, A., Zhao, H. (2015). Word polarity disambiguation using Bayesian model and opinion-level features. *Cognitive Computation*, 7 (3) 369–380.

Xu, J. M., Zhu, X., & Bellmore, A. (2012). Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 10). ACM.

Yang, Y.-H., Bhikshu, H., & Tsaih, R. H. (2011). The topic analysis of hospice care research using co-word analysis and GHSOM. *Intelligent Computing and Information Science*, 459-465.

Zhong, X., Sun, A., & Cambria, E. (2017). Time expression analysis and recognition using syntactic token types and general heuristic rules. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Volume 1, pp. 420-429.

Zhu, T., & Yu, J. (2014). A prerecognition model for hot topic discovery based on microblogging data. *The Scientific World Journal*, Volume 2014, Article ID 360934, 1-11.

Bibliographic information of this paper for citing:

Sohrabi, Babak; Raeesi Vanani, Iman, & NamAvar, Meysam (2019). "Investigation of trends and analysis of hidden new patterns in prominent news agencies of Iran using data mining and text mining algorithms." *Webology*, 16(1), Article 182. Available at: http://www.webology.org/2019/v16n1/a182.pdf