International Journal of Information Technology and Language Studies (IJITLS)

Mining Text in News Channels: A Case Study from Facebook

Said A. Salloum [1, 2], **Mostafa Al-Emran** [3], and **Khaled Shaalan** [1]
Salloum78@live.com; malemran@buc.edu.om; Khaled.shaalan@buid.ac.ae

[1] Faculty of Engineering & IT, The British University in Dubai, Dubai, UAE.
[2] University of Fujairah, Fujairah, UAE.
[3] Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Pahang, Malaysia.

Abstract. Recently, the usage of social media websites has become an attractive phenomenon in our daily life. These sites allow their users to communicate with each other through various tools. This results in learning and sharing of valuable information among their users. The nature of such information is categorized as unstructured and fuzzy text. The present study aims at analyzing textual data from Facebook and attempts to find interesting knowledge from such data and represent that knowledge in different forms. 33815 posts from 16 news channels pages over Facebook were extracted and analyzed. Findings revealed that there is a strong relationship between the Guardian and the Independent online news channels. Results indicated that there are four clusters in the study. Moreover, results showed that the overall collected data concentrated on three main topics: "*Rio de Janeiro*", "*USA elections*", and "*UK leaves the European Union*". These three main topics are considered as the hot topics that were discussed across all news channels provided by Facebook posts. Moreover, results depicted how the Text Parsing node can be employed to recognize terms and their examples in the dataset that involves the text. Other implications and future work are presented in the study.

Keywords: *Text Mining, Social Media, Facebook, News Channels.*

1. Introduction

Social networking websites make new routes for connecting people from various communities (Baumer et al., 2010). Facebook and Twitter are two extremely popular social media sites which have lately turned into an exceptionally renowned business apparatus (Chen et al., 2014). People have the opportunity via social networks to interact with individuals who have distinctive social and moral values. The diverse attainment of the internet has paved the path for the people to propagate their opinion in cyberspace, which personifies digital revolution. Social networking websites are undeniably altering the traditional norms of communication (Mhamdi, 2016). Opinions on the web are adequate to grab all the views, issues and hot topics worldwide (Kasture & Bhilare, 2015). Text mining is a new orientation of artificial intelligence showing promising developments in the text mining technology, and its application domain is also expanding with time. These websites are powerful platforms for correspondence among people that results in learning and sharing of important information (Sorensen, 2009). The most prominent social media sites are Facebook, LinkedIn, and Twitter where individuals can chat with each other through joining distinctive groups and discussion forums. These websites facilitate knowledge creation and sharing among their users by discussing different topics in various disciplines.

Text mining seems to grasp the complete automatic natural language processing. For instance, investigation of linkage structures, references in academic writing and hyperlinks in the Web writing are important sources of data that lie outside the conventional area of NLP. NLP is one of the topics that is concerned with the interrelation among the vast amount of unstructured text on social media (Salloum et al., 2016), besides the analysis and interpretation of human-being languages (Al Emran & Shaalan, 2014; Al-Emran et al., 2015).

The study is organized as follows: Section 2 provides a brief background. Other related studies are addressed by Section 3. Section 4 discusses the research methodology. The findings of this study are presented in Section 5. Conclusion and future perspectives are discussed in section 6.

2. Background

2.1. Text Mining

Salloum et al. (2017) stated that text mining has become one of the trendy fields that has been incorporated in several research areas such as computational linguistics, Information Retrieval (IR) and data mining. Text mining or knowledge discovery in the textual database of which new intriguing knowledge is characterized as the procedure of extracting unknown, easy to understand, potential and practical patterns or knowledge from the collection of large and unstructured text data or corpus (Tan, 1999; Hearst, 1999; Feldman & Dagan, 1995). Text mining is a branch of data mining that has a potentially higher business value than data mining because 80% of an organization's data is in textual format (Grimes, 2008). Nonetheless, text mining is more challenging due to the unstructured text data. Text mining is an auspicious method which extracts useful knowledge (Rajman & Besançon, 1998) by using unstructured text data. Natural language processing, data mining, and information visualization are some complicated techniques which allow and facilitate the extraction of new knowledge. In order to increase the efficiency of analyzing text, several techniques are introduced which include keyword extraction, document clustering, automatic document summarization, sentiment analysis, and topic detection and tracking.

2.2. Text Mining vs. Data mining

Text mining is basically a sub-part of data mining (Navathe & Ramez, 2000). Data mining attempts to discover interesting patterns from massive databases. Text mining, intelligent text analysis, text data mining, or Knowledge-Discovery in Text (KDT), are the names given to the procedure of extracting interesting and significant data and information from unstructured text. It is a relatively new interdisciplinary field that interrelates with other fields like data recovery, information mining, machine learning, statistics and computational linguistics. As more than 80% of data is stored in the form of text, text mining is expected to have a high business potential value (Gupta & Lehal, 2009). Knowledge might be found from numerous sources of information. However, unstructured texts remain the biggest readily available source of information (Wakade et al., 2012).

2.3. Text Mining Methods and Techniques

2.3.1 Classification of Text Mining

Although text mining is in a state of emergence, the research work involving this field is on significant volume with diverse application areas. As per these application areas, text mining can be classified as text categorization, text clustering, association rule extraction and trend analysis. In text mining, one field holds ample significance because of its efficient and promising results i-e text clustering. This unsupervised process allows objects to be classified into groups without involving any predefined categories. Relevant documents in text mining are required to show more similarities to each other than to non-relevant ones. Clustering technique is a reliable method that is usually used in analyzing an enormous amount of data such as data mining, image segmentation, document retrieval and pattern classification. The desired results include increment in precision level and recall rate of information retrieval system (Cutting et al., 1992). Clifton and Cooley (1999) claimed that text clustering has proved to be a competent tool for text theme analysis and provides a topic analysis method; groups of named entities that often occurred together. This results in performing the clustering process of the named entities grouped by the frequent item sets with the application of hyper graph-based method (Han et al., 1997). Each cluster tends to show a set of named entities and corresponds to an ongoing topic in the corpus. Topic tracking of dynamic text data has also attracted the digital arena and proved it as an interesting subject of text clustering. Montes-y-Gómez et al. (2001) proposed a text mining technique to obtain topics from online news and analyze the domination and influence of the peak news topics in comparison with other existing issues. A common experience in news reports is the effect of peak news topic (i.e. a topic with a short-term peak of frequency). On the other news topics, this term Ephemeral association was used for such influence.

2.3.2 K-Means Algorithms

There are various ways to reach desired results with clustering of data points through a particular algorithm (known as k-means). K-means is one of the most popular techniques that is used in the field. This algorithm carries out the division of ***n*** data points into ***k*** clusters minimizing the distance between each data point and its cluster's (Zaza & Al-Emran, 2015). On initial level, k-means chooses k random points from the data space, without any constraint of points in the data, eventually assigning them as centroids. Subsequent action will be the assignment of data point to the closest centroid, ultimately creating ***k*** clusters. After this first step, the centroids are reassigned so the distance between them and the points in their cluster can be reduced. Each data point is reassigned to the closest centroid.

3. Related Work

Chan et al. (2014) make it evident that social media data is vulnerable to exploitation. The proposal involves the application of structured approach and carrying out a comprehensive analysis of social media comments and a statistical cluster analysis so that the inter-relationships can be determined amid significant factors. These proposals provide assistance to quantify the qualitative social media data and cluster them on the basis of their similar features, using them for later applications such as decision-making. The data attainment process was carried out by using SAMSUNG Mobile Facebook page where Samsung smartphones were introduced. Data is primarily the comments posted by Facebook users on the captioned Facebook page. NCapture (for NVivo 10) is used for this process. There are about 128371 comments downloaded in a period of 3 months. The analysis process was applied to English comments only. Then, the content analysis included conceptual analysis and eventually, the relational analysis was used for the statistical cluster analysis. Therefore, statistical cluster analysis was used to amalgamate social media data based on the output of the conceptual analysis. Such stratagem makes it possible for the scholars to classify larger datasets into a smaller number of subsets; sometimes known as objects.

Rahman (2012) stated that the social data is used for the purpose of exploring mine intellectual knowledge and is bestowed as a systematical data mining architecture. Moreover, the primary text data comes from Facebook. In addition to that, the author accentuates information "about me" from my wall post, "age" and "comments" from Facebook, working as raw data, which then is implemented to observe the analytical approaches. Nevertheless, the author distinguished the intellectual levels for understanding fellow humans and responsibilities of the job, analyzing images, approaching to decision making and for the advertisement of their products. The extortion of intellectual knowledge from social data is preceded by using various methods involved in data mining. In particular, it delegates activities and mandatory information in which users are ascribed with respect to their peers, categories, and society connected with social networking site (e.g. Facebook).

Various tools were generated with the intention of exploring social media based on social media data analytics. Therefore, an architecture was advocated, specially designed for exploring information from Facebook by Perwitasari et al. (2015) who modified Rahman's architecture (Rahman, 2012). This architecture embraces four blocks of new units, in contrast to Rahman's architecture, together with (1) Data Collection and Temporary Storage Unit, (2) Data and Text Pre-processing Unit, (3) Network Analysis and Data Mining Unit, and (4) Knowledge Representation Unit. On the other hand, there are three main architectural features of software architecture which involve conceptual integrity, correctness, and completeness and build ability. The software was then created, which executes the structural features extracted information functionally, especially from social networking sites like Facebook, Twitter, and Instagram. The actual software illustrates four new classes (of 12 classes) which are originated from the genuine classes. Thus, it was observed that various other social media sites engender nominal attempts because of the establishment of the architectural software with numerous aspects. To reduce these attempts, the software of factory pattern method was implemented to promote structural configurability and structural flexibility.

Utilization of text mining for sentiment classification was the primary focus of (Akaichi et al., 2013). During the "Arabic Spring" period, the illustration is accomplished on Tunisian consumers' statuses on Facebook posts. The chief motive behind this is to attain valuable data regarding utilizer's sentiments as well as behaviors at this sensitive and critical period. So as to accomplish that purpose, a method founded on Support Vector Machine (SVM) and Naïve Bayes is offered. A sentiment lexicon built on the emoticons, interjections along with the acronyms', from extracted statuses updates, was established as well. Furthermore, various comparative experiments were amid two machine learning algorithms, SVM and Naïve Bayes by a training replica used for sentiment classification. The nationality of the user as well the

duration of posting their post updates that are known as Tunisian revolution, which is centralized by the uniqueness of this collection. There is no doubt that this period is incredibly special and unusual for them. Hence, the authors come up with motivating and distinct wall posts, ultimately permitting to analyses. The authors led a comparative experimental process between the Naïve Bayes and the SVM algorithms by merging different feature extractors utilizing the most renowned machine learning algorithms. Those algorithms can attain precision for classifying sentiment through combining different features. News, surveys, Facebook statuses have unique characteristics as compared to other corpuses even if the machine learning algorithms are exposed to categorize statuses with analogous performance.

The execution of correlation, clustering, and association analyses to social media have been discussed by (Mosley Jr, 2012). This is evaluated through the examination of insurance Twitter posts. It becomes easier to recognize keywords and theories in the social media data and may have the ability to facilitate the application of this information by insurers. Furthermore, insurers will proactively address potential market and client queries with efficacy after analyzing this information and implement the outcomes of the analysis in appropriate fields. As part of this evaluation, there are overall 68,370 tweets that were utilized. There are two extra kinds of evaluation to be applied on the data. The clustering analysis is the first one, which will combine the tweets depending on their similarities or dissimilarities. The second one is an Association Analysis that explores the occurrence of particular composed words.

We can observe from the existing literature that a lot of issues in Facebook were not yet analyzed and explored; one of them is the news channels' pages. As a result, we are interested to concentrate on this issue and try to build a new knowledge from these channels text analysis results. Based on the existing literature, we are seeking to answer the following research questions:

Q1: What are the most similar topics among the news channels under study?

Q2: How are the news channels documents relevant to each other?

Q3: How is the data categorized by terms?

4. Research Methodology

The datasets have been collected via Facepager software which is usually used to extract public existing data from Twitter, Facebook, and other social media based API. It collects URLs from query setup. Then, the extracted data will be stored in a local database and could be exported to a CSV format. In this study, around 37551 posts were collected from the sixteen most popular news channels sites worldwide on Facebook. A typical issue with posts content information is the existence of linguistic noise. For our situation, it would be superfluous posts that are inconsequential on prevailing topics. This view includes the objects of our data collection. Missing and garbage data have been removed from dataset and the cleaned data has been uploaded into RapidMiner tool.

During importing the dataset into RapidMiner tool, the irrelevant attributes have been excluded for enhancing the performance and data quality. Missing attributes have been removed from the analysis which contains missing data in order to increase the precision. The final number of the cleaned records that have been used for the investigation was 33815. Our collected data includes some special characters and empty cells. As per the study of Atia and Shaalan (2015), we deal with these data by removing the special characters and empty cells through the use of pre-processing techniques.

The initial steps include separation of documents into tokens with each word representation, usually known as Tokenization (Verma et al., 2014). The next step carries out the transformation process of all the characters, creating a document in a lower case. The third step includes the stop words filtering, in which this operator filters English, stop words from a document eradicating equivalent stop word from the built-in stop word list. It is required that every token represents a single English word. An operator applying eliminating all tokens identical to a stop word from file provided for the process. The file is required to possess one stop word per line. The last step associated with text processing is the filter tokens by length. This operator filters tokens based on their length; we stipulate the nominal number of characters that a token is 4, and the maximal number of characters that a token is 25.

5. Results

Q1: What are the most similar topics among the news channels under study?

As per the study of Irfan et al. (2015), we applied the similarity operator among all the news channels in order to determine the most similar topics among them. Figure 1 represents all the similarity

relationships among all the news channels. We can observe from these relationships that the most similar topics are existing between (14 and 16) news channels which stand for the Guardian and the independent online respectively. Figure 4 illustrates the distribution of clusters among all news channels.

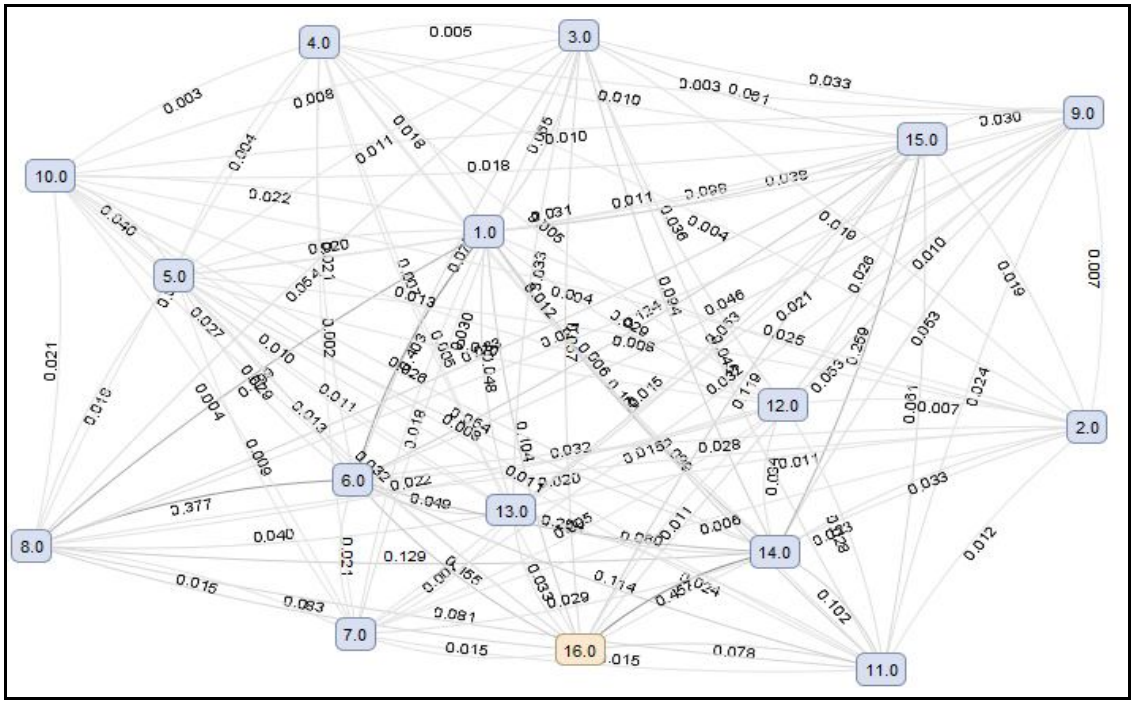

Figure 1. Similarity diagram.

Q2: How are the news channels documents relevant to each other?

According to Zaza and Al-Emran (2015), we have applied the clustering technique. We used the *k*-means algorithm through the use of different *k* values. Eventually, (*k*=4) was the most reasonable value for answering the above question. As per (Figure 2), there are four clusters. Cluster 0 contains 6 items (i.e. 6 news channels), cluster 1 includes 5 items, cluster 2 contains 1 item while cluster 3 includes 4 items. According to (Figure 3), we can notice that the clusters (0, 1, and 3) are interrelated to each other. More interesting, cluster 2 is shown to be the top cluster among the others in terms of using various keywords that were not used in other clusters (e.g. Dubai, Abu Dhabi, Sharjah, Oman, and Sohar). By analyzing the data for further investigation, we noticed that the Gulf News channel is the dominant in cluster 2. This is reasonable as this news channel is publishing news related to the Gulf region and the above mentioned examples are all real examples from this region.

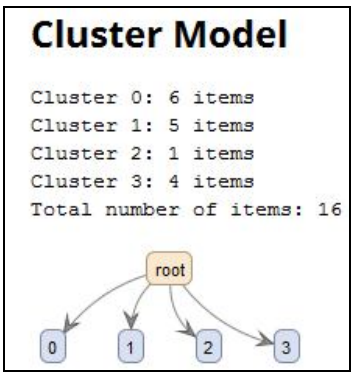

Figure 2. Cluster Model.

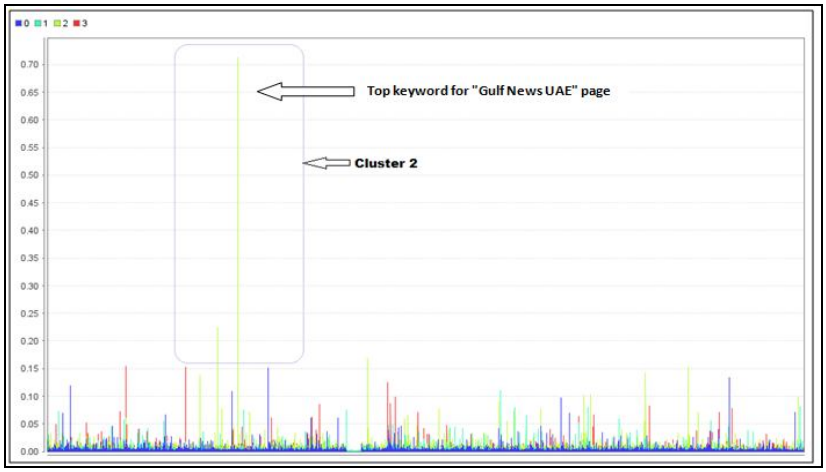

Figure 3. News channels clusters.

Row No.	id	label	cluster
1	1	ABC News	cluster_0
4	4	Associated P...	cluster_0
6	6	cnn	cluster_0
8	8	Fox News	cluster_0
10	10	NBC	cluster_0
11	11	Reuters	cluster_0
2	2	Africa 24 News	cluster_1
3	3	Aljazeera	cluster_1
7	7	Dawndotcom	cluster_1
12	12	Rio Times	cluster_1
13	13	The Canadia...	cluster_1
9	9	Gulf News UAE	cluster_2
5	5	bbc	cluster_3
14	14	The Guardian	cluster_3
15	15	The Guardian...	cluster_3
16	16	The Indepen...	cluster_3

Figure 4. News channels distribution among clusters.

Furthermore, Using the Test Parsing node, documents can be parsed in order to obtain more extensive information pertaining to phrases, terms, and rest of the entities within the category. Using the Text Cluster node, one can cluster documents into significant classes and present the concepts identified in the clusters. A group of documents can be parsed through the Text Parsing node so as to obtain the quantitative information regarding the phrases included within these documents. One is able to examine the group of documents using the Text Topic node as it creates instant relationship amongst the terms and documents, as per identified and user-defined themes. Topics signify the group of terms that explain and depict a key theme or idea. The main purpose of generating list of topics is that one can create word combinations that require analysis. When a person is able to integrate individual terms into topics, their text mining analysis can be enhanced. Integration allows narrowing the amount of text that needs to be analysed into particular categories of words that one is concerned about. An organized series of rules are created by the Text Rule Builder using small subgroups of terms that are on their whole valuable in explaining and forecasting a target variable. Every rule in the set is linked to a particular target group that includes a conjunction which suggests the existence or lack of a single or small subgroup of terms. Figure 5 indicated that the overall collected data are focused on 3 main topics. The first topic is concerned with the issue of "*Rio de Janeiro*" that reflects the 2016 Summer Olympic that were held in Brazil. The second topic is related to the issue of "*USA elections*". The third topic is relevant to the issue of "*UK leaving the European Union*". These three main topics are considered as the hot topics that were discussed across all news channels provided by Facebook posts.

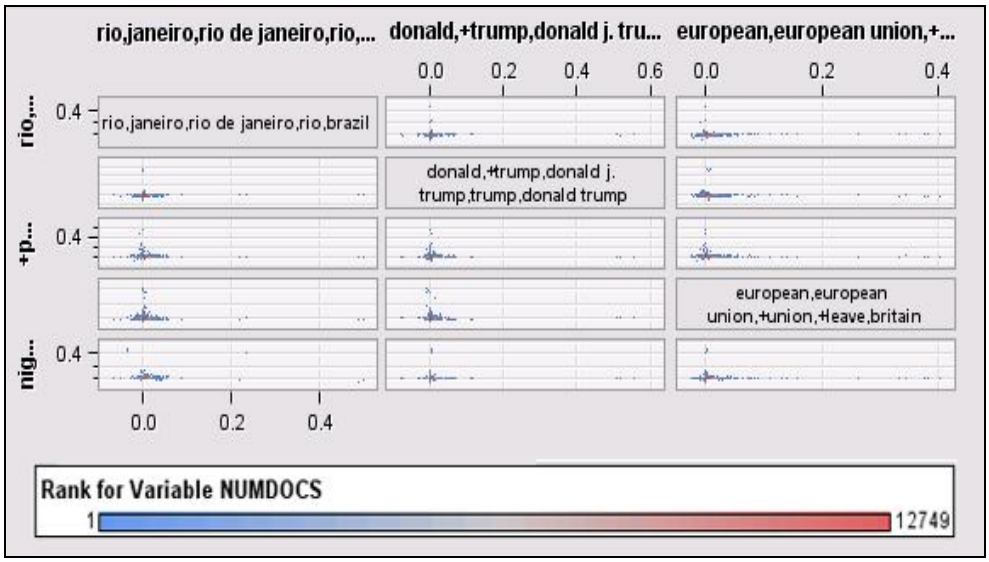

Figure 5. Topic Terms.

Q3: How is the data categorized by terms?

In the illustrated example by (Figure 6), it is depicted that how the Text Parsing node can be employed to recognize terms and their instances in a data set that involves text. Figure 6 presents the terms and their examples within the news channels sources of data. The Role shows that the Noun has the largest frequency rate as comparing with other roles presented in the group of documents.

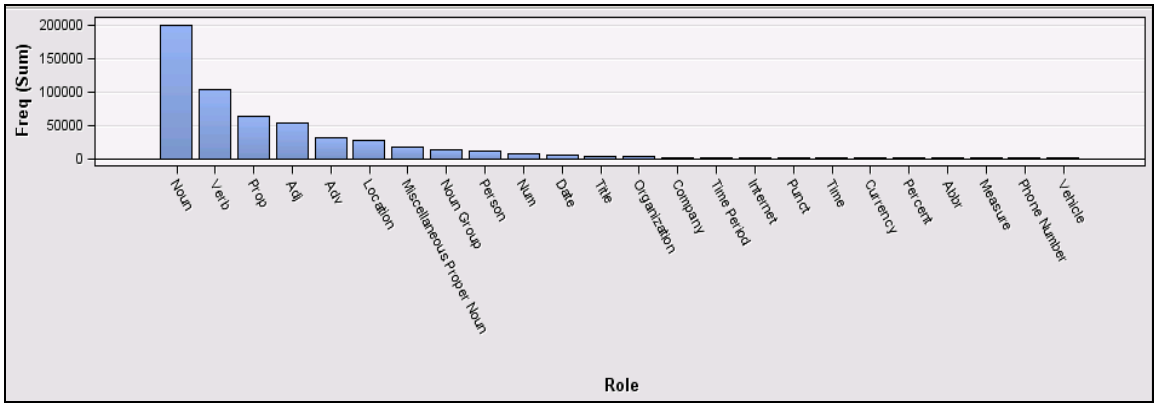

Figure 6. Role by frequency.

6. Conclusion

The evolving social media has radically transformed the communication method of people. This issue provides assistance to new agencies, companies affiliated with business arena. Specification of keywords and phrases may provide fruitful data to map out their future plans. Easy decision making helps managing their business and communication becomes facile with current and potential customers. In this research, a comprehensive explanation about text mining and its research status is provided with reliable references. We found that news channels pages on social media did not get any attention from other scholars; the reason that makes us interested in investigating this particular issue. Through the use of Facepager tool, we collected 37551 posts from Facebook which were extracted from 16 international news channels. After pre-processing, the total number of posts became 33815. Different text mining techniques have been applied on the collected data. Each of which has its own results and implications.

These methods were used in the literature. However, none of these methods were applied on news channels pages provided by Facebook.

According to Irfan et al. (2015), we used the similarity operator among all the news channels for determining the most similar topics among them. It has been concluded that there is a strong relationship between the Guardian and the independent online news channels. In relation with (Zaza & Al-Emran, 2015), we applied the clustering technique. Results revealed that there are four clusters. According to (Figure 3), we can notice that the clusters (0, 1, and 3) are interrelated to each other. More promising, cluster 2 is shown to be the top cluster among the others in terms of using various keywords that were not employed in other clusters (e.g. Dubai, Abu Dhabi, Sharjah, Oman, and Sohar). By analyzing the data for further investigation, we noticed that the Gulf News channal is the dominant in cluster 2. This is reasonable as this news channel is publishing news related to the Gulf region and the above mentioned examples are all real examples from this region. Moreover, the results showed that the overall collected data concentrated on 3 main topics: "*Rio de Janeiro*", "*USA elections*", and "*UK leaves the European Union*". These three main topics are considered as the hot topics that were discussed across all news channels provided by Facebook posts. Moreover, results depicted how the Text Parsing node can be employed to recognize terms and their examples in the dataset that involves the text. Figure 6 presented the terms and their examples within the news channels sources of data. The Role shows that the Noun has the largest frequency rate as compared with other roles presented in the group of documents. For future work, we are highly interested to collect and extract unstructured text from different news channels that provide news in Arabic language from Facebook and Twitter.

References

Akaichi, J., Dhouioui, Z., & Pérez, M. J. L. H. (2013, October). Text mining facebook status updates for sentiment classification. In *System Theory, Control and Computing (ICSTCC), 2013 17th International Conference* (pp. 640-645). IEEE.

Al Emran, M., & Shaalan, K. (2014, September). A Survey of Intelligent Language Tutoring Systems. In *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on* (pp. 393-399). IEEE.

Al-Emran, M., Zaza, S., & Shaalan, K. (2015, May). Parsing modern standard Arabic using Treebank resources. In *Information and Communication Technology Research (ICTRC), 2015 International Conference on* (pp. 80-83). IEEE.

Atia, S., & Shaalan, K. (2015, April). Increasing the Accuracy of Opinion Mining in Arabic. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)* (pp. 106-113). IEEE.

Baumer, E. P., Sinclair, J., & Tomlinson, B. (2010, April). America is like Metamucil: fostering critical and creative thinking about metaphor in political blogs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1437-1446). ACM.

Chen, X., Vorvoreanu, M., &Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, *7*(3), 246-259.

Clifton, C., & Cooley, R. (1999, September). TopCat: Data mining for topic identification in a text corpus. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 174-183). Springer Berlin Heidelberg.

Cutting, D. R., Karger, D. R., Pedersen, J. O., &Tukey, J. W. (1992, June). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 318-329). ACM.

Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In *KDD* (Vol. 95, pp. 112-117).

Grimes, S. (2008). Unstructured data and the 80 percent rule. *CarabridgeBridgepoints*.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

Han, E. H., Karypis, G., Kumar, V., & Mobasher, B. (1997, May). Clustering based on association rule hypergraphs. In *DMKD* (p. 0).

Hearst, M. A. (1999, June). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10). Association for Computational Linguistics.

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., ... & Tziritas, N. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(02), 157-170.

Kasture, N. R., & Bhilare, P. B. (2015, February). An Approach for Sentiment analysis on social networking sites. In *Computing Communication Control and Automation (ICCUBEA), 2015 International Conference on* (pp. 390-395). IEEE.

Mhamdi, C. (2016). Transgressing Media Boundaries: News Creation and Dissemination in a Globalized World. *Mediterranean Journal of Social Sciences*, *7*(5), 272.

Montes-y-Gómez, M., Gelbukh, A., &López-López, A. (2001). Discovering Ephemeral Associations among news topics. In *17th International Joint Conference on Artificial Intelligence IJCAI-01, Workshop on Adaptive Text Mining*.

Mosley Jr, R. C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. In *Casualty Actuarial Society E-Forum, Winter 2012 Volume 2* (p. 1).

Navathe, S. B., & Ramez, E. (2000). Data warehousing and data mining. *Fundamentals of Database Systems*, 841-872.

Perwitasari, A., Akbar, S., &Saptawati, G. P. (2015, November). Software architecture for social media data analytics. In *2015 International Conference on Data and Software Engineering (ICoDSE)* (pp. 208-213). IEEE.

Rahman, M. M. (2012). Mining social data to extract intellectual knowledge. *arXiv preprint arXiv:1209.5345*.

Rajman, M., & Besançon, R. (1998). Text mining: natural language techniques and text mining applications. In *Data mining and reverse engineering* (pp. 50-64). Springer US.

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2016). A Survey of Lexical Functional Grammar in the Arabic Context. *Int. J. Com. Net. Tech*, *4*(3).

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives. *Advances in Science, Technology and Engineering Systems Journal*.

Sorensen, L. (2009, May). User managed trust in social networking-Comparing Facebook, MySpace and Linkedin. In *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on* (pp. 427-431). IEEE.

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (Vol. 8, pp. 65-70).

Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and Filtering Process in Rapid Miner. *International Journal of Applied Information Systems*, *7*(2), 16-18.

Wakade, S., Shekar, C., Liszka, K. J., & Chan, C. C. (2012, January). Text mining for sentiment analysis of Twitter data. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Zaza, S., & Al-Emran, M. (2015, October). Mining and Exploration of Credit Cards Data in UAE. In *2015 Fifth International Conference on e-Learning (econf)* (pp. 275-279). IEEE.