Topic Extraction in Social Media

Ahmed Rafea, Nada A. Mostafa

Computer Science and Engineering Department
School of Science and Engineering
American University in Cairo
rafea@aucegypt.edu nadaaym@aucegypt.edu

Abstract— **Social networks have become the most important source of news and people's feedback and opinion about almost every daily topic. With this massive amount of information over the web from different social networks like Twitter, Facebook, Blogs, etc, there has to be an automatic tool that can determine the topics that people are talking about and what are there sentiments about these topics. The goal of the research described in this paper was to develop a prototype that can "feel" the pulse of the Arabic users with regards to a certain hot topic. Our experience in extracting Arabic hot topics from Twitter is presented in this paper. The unigram words that occurred more than 20 times in the whole corpus were used as features for clustering the tweets using bisecting k-mean clustering algorithm. This has resulted in purity of 0.704 and entropy of 0.275. The score generated for the quality of the generated topic was 72.5%.**

***Keywords*: Social Media Applications, Information Extraction, Clustering, Twitter**

I. INTRODUCTION

Social networks have become the most important source of news and people's feedback and opinion about almost every daily topic. Analyzing data from social networks is the key to know what people think or up to about certain topics. One of the main tasks that interest many researchers is to extract hot topics from social media.

With this massive amount of information over the web from different social networks like Twitter, Facebook, Blogs, etc, there has to be an automatic tool that can determine what people are talking about in certain locations and over certain period of time. Several researches have been done with different approaches. A lot of them achieved good performance regarding topic detection, which is basically grouping (clustering) similar data together indicating they are about the same topic. Topic extraction can be considered a next step whose goal is to label these groups through extracting a topic title for every group of data.

The research described in this paper is part of a wider objective to develop a prototype that can "feel" the pulse of the Arabic users with regards to a certain hot topic. The work related to different approaches of topic extraction using clustering and cluster labels extraction is described in the second section. The experiments conducted to decide on the features to be used for clustering, the number of clusters, and the method for topic extraction are presented in the third, fourth and fifth sections. The conclusion and future work are given in the fifth section.

II. RELATED WORK

The idea of this research domain has originated back in the 1990's with a project called TDT (Topic Detection and Tracking). The basic idea was originated in 1996 when the Defense Advanced Research Projects Agency (DARPA) realized that there is a technological need to determine the topical structure of news streams without human intervention [1]. Topic detection is the problem of identifying stories in several continuous news streams that pertain to new or previously unidentified events. Detecting the occurrence of a new event such as a plane crash, a murder, a jury trial result, or a political scandal in a stream of news stories from multiple sources is an example of topic detection [3]. The task of topic extraction can be achieved by clustering a group of news items, blogs or tweets and then discovering the labels of these clusters. These clusters labels are actually the topics extracted from this group of news items, blogs, or tweets. The work related to clustering and labeling the clusters are described in the following subsections.

A. Clustering

Hierarchical clustering was used for text clustering for topic detection [16] while agglomerative clustering with time decay was used to identify events in news [5]. The agglomerative hierarchical clustering algorithm based on the average link method for online topic detection and tracking of financial news were improved in [4]. The hierarchical agglomerative clustering technique was used for text hierarchical topic identification algorithm based on the dynamic diverse thresholds clustering [22]. The hierarchical agglomerative method was used to group articles into clusters of same events [7]. Their work aimed at extracting hot events from news feeds. The agglomerative clustering was used for topic extraction from blog entries within a neighborhood [12].The bisecting k-mean algorithm was used for topic detection by clustering key words of documents [20]. Using incremental clustering for automatic topic detection was discussed in [23] and a new topic detection method called TPIC that adds the aging nature of topics to pre-cluster stories was proposed.

978-1-4673-6404-1/13/$31.00 ©2013 IEEE

B. *Labeling Approaches*

In order to extract the topic described by a cluster, key phrase extraction techniques are mainly used to do that. Generally key-phrase extraction techniques can be categorized into simple statistics, linguistic, and machine learning.

The statistical language model was used to extract key phrases in the work presented in [17] while heuristic rules and statistics were used to develop a key phrase extraction system called KP-Miner [6]. Term Frequency * Proportional Document Frequency (TF*PDF) algorithm is used to recognize the terms that try to explain the main topics [9]

This algorithm is designed to assign heavy term weight to these kinds of terms and thus reveal the main topics. Temporal and social terms evaluation was tackled for extracting emerging topics on Twitter [2].

Part of speech tagging, formatting, and position of words were used as features to identify the terms describing the topic and high matching against the annotated data were achieved [8]. Semantic information for automatic key phrase extraction was used in [19]. An automatic online news topic key phrase extraction system was developed in which TDT algorithm was combined with aging theory for topic detection and tracking [18]. Automatic titling of electronic document using noun phrase extraction was investigated in [10].

A tool for key-phrase extraction called KEA was developed [21]. A machine- learning algorithm was used to predict which candidates are good key phrases. A tool called “Verbatim” available online was built to automatically extract quotes and topics from news feeds [13]. Support vector machines SVM, and Rocchio classifier were used in this tool.

III. FEATURES SELECTION EXPERIMENTATION

There are many features that can be used to represent the tweets such as TF-IDF (term frequency-inverse document frequency), n-grams, part of speech tags n-gram, stylistic features and/or others. Determining the features that will produce the best possible purity using the common simple representations used in clustering namely TF-IDF and n-grams is required.

The experiment steps conducted were:

1. 110 tweets over a span of 4 days were collected.
2. The tweets are manually annotated so we can get the topic of sentiment beforehand. We have 12 topic; they are all around the impact of Jan 25th revolution in Egypt. The topics are the hottest events happened during that period of time.
3. CLUTO tool for data pre-processing and clustering was used [24]. The pre-processing consists of these tasks: tokenize words, stem words, remove CLUTO English stop words translated to Arabic, calculate term weight, and represent the items to be clustered. In this experiment we did not apply stemming as we found difficulty to integrate an Arabic stemmer with the Perl script suggested by CLUTO tool to perform pre-processing. There was a problem in using Arabic letters, so we had to modify it so it can accept Arabic letters. We developed another small script to remove the stop words only; it was easier that way rather than merging both in one script. The stop words list was obtained by simply translating English list suggested by the tool as there was no Arabic stop word list available.
4. The features used in the experiments were TF-IDF, unigram, bigram, and trigram, occurred more than 20 times in the corpus, and the clustering algorithm was the bisecting k-mean with k=20.

In addition to purity measure the entropy, intra-similarity, and inter-similarity metrics were generated by the CLUTO tool. The results of this experiment are shown in the following table:

TABLE I

RESULTS OF FEATURE SELECTION EXPERIMENT

	Purity	Entropy	Intra-similarity	Inter-similarity
TF-IDF	0.573	0.385	0.227	0.004
Unigram	**0.704**	**0.275**	**0.452**	**0.008**
Unigram+ Bigram	0.694	0.289	0.452	0.008
Unigram+ Bigram+ Trigram	0.694	0.289	0.452	0.008

The four metrics shown in the table are consistent. The highest purity, the lowest entropy, the maximum intra-similarity, and minimum inter-similarity coincide when the unigram feature is used. The metrics did not decrease significantly when the combination of unigram, bigram and trigram was used. The improvements we got for purity, entropy, intra-similarity, and inter-similarity of unigram of words over the TF-IDF representations are 22.86%, 40.00%, 99.12%, and 100%. Therefore the unigram of words will be used for representing the features of tweets.

IV. EXPERIMENT TO DETERMINE THE NUMBER OF CLUSTERS

Determining the number of clusters is important in bisecting k-mean and for other clustering algorithms as well.

Our approach was to perform 3 experiments that differ from each other in the numbers of clusters. We have 12 predetermined labels for our clusters, so we decided to perform experiments on 6, 12 and 20 way clustering to compare results.

The experiment steps conducted were:

1. The 110 tweets preprocessed and represented in TF-IDF used in the feature selection experiment were used in this experiment
2. Run the bisecting k-mean algorithm with k= 6,12, and 20 on the preprocessed tweets

The results obtained are shown in the following table:

TABLE II

RESULTS OF DETERMINING NUMBER OF CLUSTERS

K	Purity	Entropy	Intra-similarity	Inter-similarity
6	0.391	0.674	0.081	0.003
12	0.473	0.515	0.136	0.004
20	**0.573**	**0.385**	**0.227**	**0.004**

The best purity, entropy, and intra-similarity, were obtained when k=20. The inter-similarity was a little bit worse when the number of clusters increased. These results were expected but more experimentation needs to be conducted on a bigger corpus. The results for 12 clusters were 0.515 for entropy and 0.473 for purity. In general, the clustering quality improved when the number of cluster increased. But for sure it could give more than one cluster containing tweets related to the same topic. However in real application we do not know the actual number of topics of the retrieved tweets. So for the online clustering to extract topics of a set of 100 tweets, we will use k=20.

V. EXPERIMENT TO DETERMINE THE CLUSTERS LABELS

Key phrase extraction is the widely used technique for labeling the cluster, which, actually can be considered the topic discussed in the cluster of tweets. One of the key phrase extraction techniques proved to be accurate is the KP-Miner [16]. Finding out whether KP-miner could be used for determining the clusters labels was investigated. The experiment steps conducted were:

1. The 110 tweets pre-processed, manually annotated with the sentiment topic, and represented in unigram features used in the feature selection experiment were used in this experiment.
2. Run the bisecting k-mean algorithm with k= 20 on the pre-processed tweets
3. Select the semantic topic of the majority of the tweets in the cluster to be the topic describing the cluster
4. Run the KP-Miner on each cluster and extract the first 3 key phrases
5. The following procedure was applied to get a quantitative measure of how good the KP-miner tool did:
    a. If the first key phrase generated matched the sentiment topic phrase we gave the tool score 1.0
    b. If the first key phrase generated was part of the sentiment topic phrase we gave the tool score 0.75
    c. If the second key phrase generated was part of the sentiment topic phrase we gave the tool score 0.5
    d. If the key phrases generated was not part of the sentiment topic phrase we gave the tool score 0.0
    e. If key phrases were generated while no sentiment topic was given by the human annotator we gave the tool score 0.0

The total score of the KP-miner tool using the above procedure was 14.5 out of 20 as we have 20 clusters, which represents 72.5%.

Clustering algorithm could generate clusters with different topics of tweets inside each cluster. Comparing the key phrases generated with these topics is another measure for the KP-miner capability to label clusters generated. The best topic phrase of the tweets in a cluster is the topic annotated by the human annotator to the majority of the tweets in a cluster. The second best topic phrase is the topic that was assigned to number of tweets less than the tweets assigned the best topic. The third best and fourth best can be defined in the same way.

Applying the above steps but changing the evaluation procedure described in step 5, as follows:

a. If the first key phrase generated matched the best topic phrase of the tweets we gave the tool score 1.0
b. If the first key phrase generated matched the second best topic phrase of the tweets we gave the tool score 0.75
c. If the first key phrase generated matched the third best topic phrase of the tweets we gave the tool score 0.5
d. If the first key phrase generated matched the fourth best topic phrase of the tweets we gave the tool score 0.25

The total score of the KP-miner tool using the second procedure was 15.75 out of 20 as we have 20 clusters, which represents 78.75%.

Although the experiments were run on small number of tweets and the number of tweets per cluster was between 3 and 13 the KP-miner shows good performance as extracting the topic of a cluster is very challenging task.

VI. CONCLUSION AND FUTURE WORK

The objective of the research conducted in this paper was to decide on the features to be used for clustering Arabic tweets, the number of clusters that is very important parameter for many clustering algorithms, and the method for extracting the cluster topic.

The unigrams of words were found to be the best features.

The clustering quality resulted from applying the bisecting k-mean was 0.704 for purity and 0.275 for entropy. These results were better than using TF-IDF of words as features where the results were 0.57 for purity and 0.36 for entropy.

Therefore unigrams were decided to be the features we used for clustering. However more research is needed to investigate other clustering algorithms, study the impact of removing the whole set of stop words and perform stemming.

A methodology for evaluating Key-phrase extraction algorithm to recognize the sentiment topic contained in a cluster was developed and applied on a small set of 110 annotated tweets. The score generated for the quality of the generated topic was 72.5%. There is still work to be done here, as other methods using different approaches still need to be investigated such as considering: name entity recognition as most of the topics are actually taking about name entities, hash tags, supervised learning approach, and hybrid approaches.

The research work presented in this was part of the research to identify hot topics and classify the sentiment of users toward these topics in Tweeter [11], [15], and [16].

Acknowledgement

This work was done in a project entitled Sentiment and Opinion Mining for Arabic Web funded by ITIDA, Ministry of Communication and Information Technology (MCIT). I would like to acknowledge: Sara Morsy and AmiraShoukry who worked in this project and some parts of the material used in this paper were done with their collaborations and published in the proceedings of International conferences as indicated in the paper.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, 1998, "Topic Detection and Tracking Pilot Study Final Report" In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop pp. 194-218

[2] M. Cataldi, L. Di Caro, and C. Schifanella, 2010," Emerging topic detection on Twitter based on temporal and social terms evaluation." In Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD '10). ACM, New York, NY, USA, , Article 4 , 10 pages.

[3] Y. Choi, Y. Jung, and S. Myaeng," Identifying Controversial Issues and Their Sub-topics in News Articles" Book Title: "Intelligence and Security Informatics" Book Series Title:" Lecture Notes in Computer Science" , 2010,Chen,H. et al. (Eds.) Springer Berlin / Heidelberg pp. 140-153

[4] X. Dai, Q. Chen, X. Wang, and J. Xu , "Online topic detection and tracking of financial news based on hierarchical clustering," Machine Learning and Cybernetics (ICMLC), 2010 International Conference on , vol.6, no., pp.3341-3346, 11-14 July 2010

[5] X. Dai and Y. Sun, 2010, "Event identification within news topics," Intelligent Computing and Integrated Systems (ICISS), 2010 International Conference on , vol., no., pp.498-502, 22-24

[6] S. El-Beltagy and A. Rafea.," KP-Miner: A keyphrase extraction system for English and Arabic Documents", Information Systems (2008)

[7] Z. Huang and A. Cardenas, "Extracting Hot Events from News Feeds, Visualization, and Insights", 2009

[8] S. Jain and J. Pareek, 2009, "KeyPhrase Extraction Tool (KET) for Semantic Metadata Annotation of Learning Materials," 2009 International Conference on Signal Processing Systems , vol., no., pp.625-628, 15-17.

[9] K. Khoo and M. Ishizuka, 2002, "Topic extraction from news archive using TF*PDF algorithm," Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on , vol., no., pp. 73- 82, 12-14

[10] C. Lopez, V. Prince, and M. Roche, "Automatic titling of electronic documents with noun phrase extraction," Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of , vol., no., pp.168-171, 7-10 Dec. 2010

[11] S. Morsy and A. Rafea, 2012, "Improving Document-Level Sentiment Classification Using Contextual Valence Shifters", Natural Language Processing and Information Systems Lecture Notes in Computer Science Volume 7337, 2012, pp 253-258

[12] M. Okamoto and M. Kikuchi, 2009, "Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries." In Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology (AIRS '09), Gary Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko Aizawa, Kazuko Kuriyama, Masaharu Yoshioka, and Tetsuya Sakai (Eds.). Springer-Verlag, Berlin, Heidelberg, 181-192

[13] L. Sarmento and S. Nunes, "Automatic Extraction of Quotes and Topics from News Feeds" http://hdl.handle.net/10216/7080

[14] Y. Seo and K. Sycara, "Text Clustering for Topic Detection", Carnegie Mellon University, 2004

[15] A. Shoukry and A. Rafea, 2012, "Sentence-level Arabic sentiment analysis", Collaboration Technologies and Systems (CTS), 2012 International Conference on , vol., no., pp.546-550, 21-25 May 2012 doi: 10.1109/CTS.2012.6261103

[16] A. Shoukry and A. Rafea, 2012, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining", The Fourth Workshop on Computational Approaches to Arabic Script-based Languages. AMTA 2012 - San Diego, CA USA, November, 2012

[17] T. Tomokiyo and M. Hurst., 2003, "A language model approach to keyphrase extraction". In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18 (MWE '03), Vol. 18. Association for Computational Linguistics, Stroudsburg, PA, USA, 33-40.

[18] C. Wang, M. Zhang, L. Ru, and S. Ma, 2008, "An Automatic Online News Topic Keyphrase Extraction System," Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on, vol.1, no., pp.214-219, 9-12.

[19] X. Wang, D. Mu, and J. Fang, 2008, "Improved Automatic Keyphrase Extraction by Using Semantic Information", Conference on Intelligent Computation Technology and Automation (ICICTA), 2008 International, vol.1, no., pp.1061-1065, 20-22.

[20] C. Wartena and R. Brussee,"Topic Detection by Clustering Keywords," Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on, vol., no., pp.54-58, 1-5 Sept. 2008

[21] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning, 1999, "KEA: Practical automatic keyphrase extraction". In Proceedings of the fourth ACM conference on Digital libraries (DL '99). ACM, New York, NY, USA, 254-255.

[22] Y. Xu, G. Quan, Z. Xu, and Y. Wang, 2009, "Research on Text Hierarchical Topic Identification Algorithm Based on the Dynamic

Diverse Thresholds Clustering," Conference on Asian Language Processing, 2009. IALP '09. International, vol., no., pp.206-210.

[23] X. Zhang and Z. Li, 2010, “Automatic Topic Detection with an Incremental Clustering Algorithm” Wang, F.L. et al. (Eds.): WISM 2010, LNCS 6318, pp.344-351, Springer-Verlag

[24] CLUTO A Clustering Toolkit Release 2.1.1 George Karypis karypis@cs.umn.edu University of Minnesota, Department of Computer Science