Wiki Spots Model for Annotating Short Text Documents Using Wikipedia Hyperlinks

Eman Ismail, Walaa Gad, Khaled Bahnsy

Information Systems Department, Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

emanismail,walaagad,khaled.bahnsy@cis.asu.edu.eg

Abstract

With the rapid growth of social applications, millions of short text documents have been produced every day. Classification of short text is a significant challenge in information retrieval due to the text noise and sparseness. In this paper, Wiki_Spots model is proposed to annotate short text documents to their best label. Wiki_Spots utilizes Wikipedia link structure to enrich the documents with new features. The new representation contains all spots in the text, each spot represents one document topic. The proposed Wiki_Spots represents the documents as vector of topics. Wiki Spots refers to each spot with Wikipedia page. The proposed Wiki_Spots model exploits Naive Bayes as classifier to estimate its performance. The experimental results show that Wiki_Spots model is significant in text annotation. The performance measures recorded 83% and 84% using recall and precision, respectively.

Keywords: *Wikipedia Knowledge base, text annotation, document representation, document classification.*

1. Introduction

With the explosion of social communication, a huge numbers of short text documents become available with different forms such as, Web snippets, chat messages, forums.., etc. Therefore, it is important to process and manage them to help people to retrieve their relevant information successfully. Thus, classification of such type of documents has a significant challenge in many Information Retrieval (IR) applications.

Short text documents (STDs) are sparse and do not have sufficient words to gain good similarity measures. Moreover, STDs are noisy due to using nonstandard words and many spelling mistakes. Thus, traditional techniques do not reach a good accuracy. The Bag-of-Words (BOW) technique represents each document as independent features. Each vector consists of features weight. The weight is the number of occurrence of each feature in the document.

Recently, there are two methods to solve the problems of short text documents. The first one replaces the documents features with fewer concepts. The second one enriches the documents with additional semantics information for better context understanding. The semantic information is extracted using Wikipedia or Ontology. Wikipedia is the largest repository on the internet. It contains more than 4 million articles pages in different topics. Document is represented as bag of concepts as in [1]–[5]. Ontology [24], [25] is a hierarchal structure which describes specific domain and provides semantic information based on the

relationships between concepts. WordNet is ontology for English language. In [6], [7] WordNet is used to enrich the text with new features containing the topics covered with in text. This step improves the accuracy compared to BOW.

The reduction method reduces the dimension of the features space by representing documents with their corresponding topics. The topics are used as keywords to annotate documents [8]. In this paper, a novel model is proposed to enrich documents with semantic information using Wikipedia hyperlink. Wikipedia inherits the characteristics of the web, and combines the textual information and cross references with hyperlinks. Hyperlinks are Wikipedia spots or anchors. It links some terms and spots which are important in the document to other Wikipedia articles. These ways provide an additional information to the reader to help in understanding the text topic.

The proposed Wiki_Spots model is based on Wikipedia as background knowledge for classifying short text documents. It enriches the document terms with Wikipedia spots. Wikipedia spots include the unigrams terms and phrases. Phrases could be bigrams, trigrams..., etc. Each spot is given a weight depend on the number of occurrence in all Wikipedia articles.

The proposed Wiki_Spots is evaluated on short text "web snippets" dataset. Snippets means that the document has a few number of terms, and it is difficult for the term to appear more than once. The performance measures of Wiki_Spots model are compared with another techniques [8]–[10]. These techniques are using WordNet as lexical database. The results ensure the effectiveness the proposed model, which gained 83% in precision and 84% in recall.

The rest sections are organized as follows. In section 2, the previous annotation techniques are presented briefly. The proposed model Wiki_Spots is totally explained in section 3. In section 4, the results and analysis of the proposed Wiki_Spots are showed in details. In section 5 conclusions are introduced.

2. Literature Overview

Recently, classification of short text documents is very important in research area. The main two techniques are enrichment and reduction techniques. The enrichment techniques expand the BOW features dimension with extra data to better understand the document. On the other hand, the reduction techniques decrease the number of features as they replace the original features with the most important terms to represent the documents. Both of them solve the limitations of short text documents with different ways as follows:

2.1 Enrichment_based_Classification

Enrichment methods [5], [6], [8], [10]–[15] are utilized to enrich the short documents with extra semantic data. The enrichment extract new terms from knowledge bases such as Wikipedia. Wikipedia is used to extract the topics that are related to document terms. The extracted topics are used to be extra features added to the original terms. This way is focused on the topic and solve the problem of data sparseness.

Latent Dirichlet allocation (LDA) is used for document enrichment [6], [8], [16]. LDA [10] executes latent semantic analysis using some probabilistic models to contain the

synonym and polysemy. Moreover, it extracts the topics covered by the short text, which is exploited to enrich the traditional BOW [9] text document representation.

Wikipedia knowledge is used to apply semantic analysis in text documents using TAGME as topic annotator [10]. TAGME identifies some meaningful sequences of terms in the short text. Then, it refers to these terms with Wikipedia pages which explain their topics. These topics are used to represent the document [8]. Consequently, TAGME proposed function to rank the highest topics to annotate the short text documents.

In [10], Phan et al. present framework to build classifier that works on short text documents. The classifier was built on the discovered topics from the dataset and some labelled training data. Therefore, LDA is utilized to apply latent semantic analysis (LSA) to perform topic modelling on the input text. LDA enriches the BOW representation with topics extracted from Wikipedia knowledge.

Wikipedia based annotator approach [8] was proposed to classify short text documents. It detects the main topics in the text documents. These topics presented in the form of Wikipedia pages. Then, a novel classification algorithm is proposed to calculate the similarity between documents and their categories or labels. Then, the extracted topics of text documents and the Wikipedia structure are exploited to measure the similarity. This approach does not expand the feature space with the extracted topics. But, it organizes them into group of top topics. These topics are chosen with the help of TAGME and based on ranking function. Moreover, it exploits TF-IDF formula [9] to exchange terms with topics and document with categories in order to apply topic based enrichment representation.

Furthermore, Classification Based Enrichment Representation (CBER) model was proposed [17] to classify short text documents to their best labels. It does not expand the feature dimension. It extracts the concepts and gets the relationships between concepts to solve the disambiguation problems. It enriches the terms with semantic weight, so the terms that are defined on WordNet gain more weight. The semantic weight is calculated based on the relationships between the desired concept and the other concepts in the same documents.

In [18], Classification approach was proposed to classify the short messages "Tweets" using Wikipedia knowledge. The method enriches the original text (message) with extra text from Wikipedia articles and maps each message to their corresponding Wikipedia page. It exploits the distance between their Wikipedia pages to measure the similarity between messages.

2.2 Reduction_based_Classification

LDA algorithm is utilized to weight the topics that extracted from the document with different weights [11], [19]. These topics are used to represent the document rather than the document features. Then, a semantic relationship is calculated between the extracted topics and the document words.

In [7], [13], authors reduce the document features using WordNet. WordNet is used to replace the document features with more semantic data for better classification accuracy. Soucy and Mineau [9] followed the same technique of extracting new terms from WordNet. Then, the terms that their weights are greater than a specific threshold are selected related to a weighting function in [9].

An algorithm was proposed to collect the results returned by any search engine [6]. Then, it groups them into clusters, which are annotated with meaningful phrases. These phrases describe the topics covered by the returned results. The results are described into graph of concepts rather than Bag of Words (BOW) using TAGME topical annotator.

In [20], it was showed a simple and scalable approach to assign label to short text document using few words called query words. These words are selected from the input text document that must represent the document content and topics.

Many drawbacks have been found in the previous methods. The enrichment methods in [14], [18], [21], [22], increase the document dimensions and increase the classification processing time due to adding extra data to the document features. In [7], [20], Wikipedia knowledge base or WordNet are used to extract the topics of the documents which treat the document as vector of topics. This way focused on the terms that are related to document topics and then neglect to ther important terms.

3. The Proposed Wiki_Spots Model

The proposed Wiki_Spots model, is showed in figure 1. Wiki_Spots model follows the enrichment techniques to overcome the limitation of short text documents of being too short. So, it is important to increase the document length with new data to represent document semantically. It enriches the short text documents with new information provided by Wikipedia. It inherits the characteristics of the web and combines the textual information and cross references with hyperlinks called spots or anchors. Moreover, it connects the spots of each document, which are the significant terms, to another Wikipedia articles. The Wiki_Spots provides the users with a quick way to access additional information. Moreover, the text spots provide more important information about the text.

The proposed Wiki_Spots model uses Wikipedia as background knowledge to classify short text documents. It enriches the document terms with Wikipedia spots, which are annotated manually by the authors of the article. Wikipedia spots include the unigrams terms and phrases. Phrases may be bigrams, trigrams..., etc. Wiki_Spots model consists of:

- Performing text cleaning on the input document using a spell checker.
- Retrieving all Wikipedia pages and spots from Wikipedia dumps.
- Mapping the text with spots with Wikipedia dumps to extracts all spots in the text.
- Linking each spot with Wikipedia page using these equation1.

Each document is represented as vector of spots.

$$\mathrm{VoteSCore}_b(p_a) = \frac{\sum_{p_b \in p(b)} \mathrm{Rel}(p_b, p_a) * \frac{p_b}{p}}{p(b)} \quad (1)$$

p_b.is one of a candidate sense (Wikipedia page) of spot a. b is a spot, which belongs to all text spots. $p(b)$is all senses of spot b. The fraction $\frac{p_b}{p}$ refers to the probability of b pointing to wiki page p_b, which calculates the numbers of times of the spot b refers top$_b$.

Rel (p_b, p_a), refers to the relatedness between two Wikipedia articles to consider the number of incoming and outgoing links of the two articles (pages).

$$\mathrm{Rel}(p_b, p_a) = \frac{\log(\mathrm{M}(|\mathrm{inp}(p_b)|, |\mathrm{inp}(p_a)|)) - \log(\mathrm{M}(|\mathrm{inp}(p_b) \cap \mathrm{inp}(p_a)|))}{\log(\mathrm{W}) - \log(\mathrm{N}(|\mathrm{inp}(p_b)|, |\mathrm{inp}(p_a)|))} \quad (2)$$

The term W refers to the size of Wikipedia articles. inp (p_b) is the number of incoming links to page p_b, N means the minimum number and M means the maximum number.

Wiki_Spots model applies spell checker algorithm to remove the spelling mistakes in short text documents. The spell checker algorithm uses special dictionary to check the validation of characters and words for given language the algorithm steps as follows:

- Parse the document to extract tokens (words) and check their corrections.
- Analyze the tokens by breaking them to their root using stemming algorithm.
- Morphological analysis to check which tokens are valid using dictionary.
- Using n-gram similarity to suggest similar correct words from the dictionary in case of incorrect words found.

The proposed Wiki_Spots model represents short text documents as vectors of spots as in equation3. Wiki_Spots classifies the text in certain predefined categories, by exploiting the annotation process of TAGME [5]. It uses Wikipedia dumps, which it is offline database version for all articles, concepts, and links of each article and also the redirect links found in Wikipedia. Wikipedia spots represent different topics about the text documents. Therefore, Wiki_Spots model exploits these spots to enrich the text to understand the topics covered by the text. The proposed model represents the documents as graph of topics. This representation is more powerful than the traditional BOW method. BOW uses the TFIDF (term frequency inverse document frequency) [9] for representation.

$$D = \begin{bmatrix} SW11 & SW12 & \dots & SW1n \\ SW21 & SW22 & \dots & SW2n \\ \vdots & \vdots & \ddots & \vdots \\ SWm1 & SWm2 & \dots & SWmn \end{bmatrix} \quad (3)$$

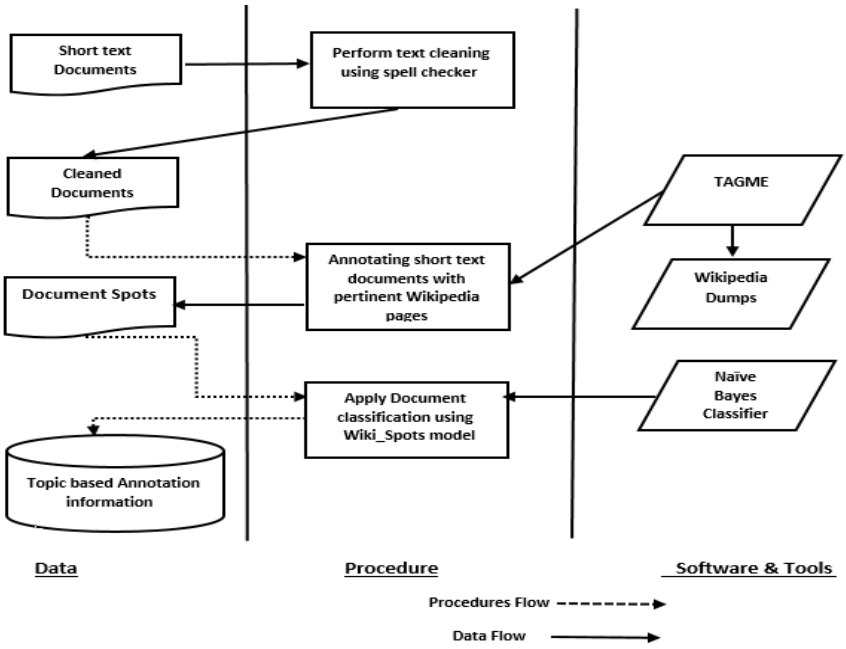

Figure 1. Wiki_Spots Model Overflow

Each spot is given a weight SW that depends on the relationships between this spot and other spots in the same documents. *n* represents the number of spots extracted from each document and *m* represents the documents number. The weight of each spot is calculated as follow in equation (4) where *s* is all document spots, *b* is one of document spots and Vote Score as in equation 1.

$$SW = \sum_{b \in s} \text{VoteSCore}_b\,(p_a) \qquad (4)$$

In the last, the Wikipedia spots vectors are used to represent the short text documents. Naive Bayes Classifier (NBC) is used for classification. NBC is very simple probabilistic classifier and high scalable.

4. Experimental Results

As in [8], [10], Google snippets dataset is used to evaluate the proposed Wiki_Spots model. The dataset contains short text documents with 13 terms on average.

4.1 Dataset Description

Google snippets dataset has been produced by Phan et al. [10]. It contains 12K snippets from Google. The snippets dataset is divided into eight classes. The class labels are business, computers, culture-arts-ent, education-science, health, politics-society, engineering and sports. Table 1 shows the number of documents per each class. Google snippets dataset contains two types of data, one for training and the other for testing

Table 1. Snippets dataset classes

Class_Label	# of Doc Per Class
C1/ Business	1200
C2/ Computers	1200
C3/ Culture-Arts-Ent.	1880
C4/ Education-Science	2360
C5/ Engineering	220
C6/ Health	880
C7/ Politics-Society	1200
C8/ Sports	1120

4.2 Classification of Short text documents (STDs)

Naive Bayes (NB) classifier is used in classification of STDs. NB is based on Bayes theorem with a strong naïve independent assumption between predictors (features) as in equation 5. Naive Bayes classifier is considered one of the highest scalable scoring and building model. It is also efficient and fast [23]. In addition, it is a good dealer with short text documents that help us getting very good results.

$$P(Cl|A) = \frac{P(A|Cl)P(Cl)}{P(A)} \qquad (5)$$

Where P (Cl|A) is the posterior probability of class Cl given predictor A. P (A|Cl) is probability of predictor A given the class Cl. P(Cl) is the prior probability of class Cl and P(A) is also the prior probability of the predictor A.

Cross fold validation is used to measure the relative absolute error, which helps in measuring the accuracy of classification over the training dataset. As shown in Figure2, the cross fold is stopped till 9 as the error decreases.

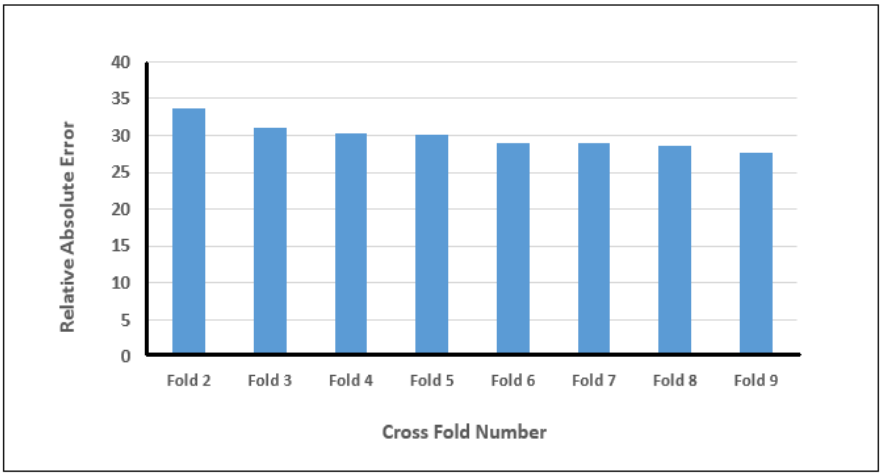

Figure 2. Wiki_Spots cross folding validation

4.3 Performance measure

Recall (Rec), Precision (Pre), F-measure (Fm) and Accuracy (Acc) are used to assess the Wiki_Spots performance.

$$\text{Rec} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}} \tag{6}$$

$$\text{Pre} = \frac{\text{Tp}}{\text{Tp} + \text{Fp}} \tag{7}$$

$$\text{Fm} = 2 * \frac{\text{Rec} * \text{Pre}}{\text{Rec} + \text{Pre}} \tag{8}$$

$$\text{Acc} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn}} \tag{9}$$

where Tp "True Positive" measures the proportion of documents that are correctly selected to their classes; Fp "False Positive" measures the proportion of documents that are incorrectly rejected from their classes; Fn "False Negative" measures the proportion of documents that are incorrectly selected to a class; and Tn "True Negative" measures the proportion of documents that are correctly rejected from a class.

4.4 Results and Analysis of Wiki_Spots Model

In table 2, Wiki_Spots model is reached 84% and 83% in terms of precision and recall respectively running the classifier on 1000 documents selected randomly from the dataset. Wiki_Spots model is also compared with topical annotator [8]. In [10] Phan et al. and traditional BOW method with SVM and MaxEnt. Phan et al. [10], they build framework to extract documents topics using latent Dirichlet Allocation (LDA) to classify short text documents (STDs) with them. Then annotates STDs to their best topics. Topical annotator [8] is based on enriching the documents representation with the extracted topics, but their results are close to Phan et al. [10]. Wiki_Spots accuracy is 84%, which outperform the Topical classifier, SVM and MaxEnt that reach accuracy 81%, 74.93% and 65.75%, respectively.

However, Wiki_Spots gets good results compared with other classifiers in [8], [10]. The experimental results showed that Wiki_Spots model has better and higher results than Word Vector Term Frequency (WVTF) in ref [17], but not higher than Classification Based Enrichment Representation (CBER) model, which reached accuracy 94%. The CBER assigns additional weighting score to document terms that are semantically related, which cause increasing the accuracy of the classification. If the documents are written in formal English language, CBER model will give efficient results.

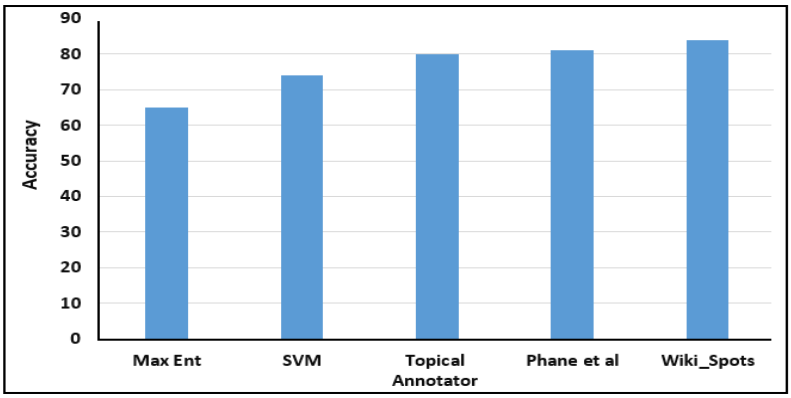

Figure 3.Evaluation for Wiki_Spots over snippet dataset

Wiki_spots model results outperform the previous models using precision, Recall and F-measure, as it reached accuracy 84% compared to Phan et al. [10], topical annotator [8] and with the traditional BOW using SVM, MaxEnt as shown in 3. The main contribution of this work is the proposed model employs Wikipedia knowledge base and TAGME to extract text spots for documents representation. Each spot links to different Wikipedia pages as possible sense and gives new weight based on the spot importance in the article.

Table 2. Wiki_Spots Results

Class_Label	TruePositive Rate	FalsePositive Rate	Precision	Recall	F_Measure
Business	o.792	0.034	0.767	0.792	0.78
Computers	0.864	0.01	0.923	0.864	0.893
Culture_arts_ent	0.88	0.047	0.728	0.88	0.797
Education	0.872	0.011	0.916	0.872	0.893
Engineering	0.768	0.031	0.78	0.768	0.774
Health	0.752	0.03	0.783	0.752	0.767
Politics-society	0.888	0.019	0.867	0.888	0.877
Sports	0.84	0.009	0.929	0.84	0.882
Weighted Avg.	0.832	0.024	0.837	0.832	0.833

5. Conclusion

In this paper, short text classification can be improved by enriching the text representation with semantic information using Wikipedia. A novel, scalable, and efficient Wiki_Spots model is proposed for classifying Short text documents. The main contribution of Wiki_Spots model is employing Wikipedia Knowledge base to identify spots in a text. Furthermore, new spots are utilized to represent the short text documents as a vector of topics. A new weighting function is proposed to reflect the importance of the spot as entity of the text. Wiki_Spots model is evaluated on snippets data set, which limits the occurrence of the terms in the same document. The experimental results show a good performance of the proposed Wiki_Spots model.

References

[1]. P. Wang and C. Domeniconi, "Building semantic kernels for text classification using wikipedia," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 713–721.

[2]. X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009, pp. 389–396.

[3]. A. Huang, D. Milne, E. Frank, and I. H. Witten, "Clustering documents using a wikipedia-based concept representation," in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2009, pp. 628–636.

[4]. P. Wang, J. Hu, H.-J. Zeng, L. Chen, and Z. Chen, "Improving text classification by using encyclopedia knowledge," in Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007, pp. 332–341.

[5]. P. Ferragina and U. Scaiella, "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010, pp. 1625–1628.

[6]. U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita, "Topical clustering of search results," in Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012, pp. 223–232.

[7]. L. H. Patil and M. Atique, "A semantic approach for effective document clustering using wordnet," arXiv preprint arXiv: 1303.0489,2013.

[8]. D. Vitale, P. Ferragina, and U. Scaiella, "Classification of short texts by deploying topical annotations," in European Conference on Information Retrieval. Springer, 2012, pp. 376–387.

[9]. P. Soucy and G. W. Mineau, "Beyond tfidf weighting for text categorization in the vector space model," in IJCAI, vol. 5, 2005, pp.1130–1135.

[10]. X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp. 91–100.

[11]. A. Bouaziz, C. Dartigues-Pallez, C. da Costa Pereira, F. Precioso, and P. Loret, "Short text classification using semantic random forest," in International Conference on Data Warehousing and Knowledge Discovery. Springer, 2014, pp. 288–299.

[12]. X. Sun, H. Wang, and Y. Yu, "Towards effective short text deep classification," in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011, pp. 1143–1144.

[13]. [13] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie, "Short text classification: A survey," Journal of Multimedia, vol. 9, no. 5, pp. 635–643, 2014.

[14]. J. Sedding and D. Kazakov, "Wordnet-based text document clustering," in proceedings of the 3rd workshop on robust methods in analysis of natural language data. Association for Computational Linguistics, 2004, pp. 104–113.

[15]. W.-T. Yih and C. Meek, "Improving similarity measures for short segments of text," in AAAI, vol. 7, no. 7, 2007, pp. 1489–1494.

[16]. J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, pp. 782–792.

[17]. E. Ismail and W. Gad, "CBER: An effective classification approach based on enrichment representation for short text documents," Journal of Intelligent Systems.

[18]. Y. Genc, Y. Sakamoto, and J. V. Nickerson, "Discovering context: classifying tweets through a semantic transform based on wikipedia," in International Conference on Foundations of Augmented Cognition. Springer, 2011, pp. 484–492.

[19]. L. Yang, C. Li, Q. Ding, and L. Li, "Combining lexical and semantic features for short text classification," Procedia Computer Science, vol. 22, pp. 78–86, 2013.

[20]. A. Sun, "Short text classification using very few words," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012, pp. 1145–1146.

[21]. C. Makris, Y. Plegas, and E. Theodoridis, "Improved text annotation with wikipedia entities," in Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM, 2013, pp. 288–295.

[22]. T. Pedersen, S. Patwardhan, and J. Michelizzi, "Wordnet: Similarity: measuring the relatedness of concepts," in Demonstration papers at HLT-NAACL 2004. Association for Computational Linguistics, 2004, pp. 38–41.

[23]. V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," International Journal of Artificial Intelligence & Applications, vol. 3, no. 2, p. 85, 2012.

[24]. Salama, S. I. A., & Mohammed, T. A. (2015). "Towards the Semantic Web: Design an Ontology and use it in Semantic Query," Egyptian Computer Science Journal, 39(1), 2015.

[25]. Gawich, M., Alfonse, M., Aref, M., & Salem, A. B. M. (2017). Developing a System for Medical Ontology Evolution. Egyptian Computer Science Journal (ISSN-1110-2586), 41(2), 2017.