Journal of Computer Science

Review

Short Text Mining: State of the Art and Research Opportunities

[1]Mohamed Grida, [2]Hasnaa Soliman and [2]Mohamed Hassan

[1]Department of Industrial Engineering, Zagazig University, Egypt
[2]Department of Information System, Zagazig University, Egypt

Article history
Received: 03-06-2019
Revised: 12-07-2019
Accepted: 11-10-2019

Corresponding author:
Mohamed Grida
Department of Industrial Engineering, Zagazig University, Egypt
Email: mogrida@zu.edu.eg

Abstract: With the growing number of connected online users producing a tremendous amount of unstructured short-texts daily, understanding and mining these data becomes very useful for individuals, governments and companies for identifying the public users' attitudes towards different entities, such as products, services, events, places, organizations and topics. However, analyzing these short-texts using traditional methods becomes a significant challenge due to the shortness and sparsity nature of short-texts. To address such challenges, the literature introduced a broad spectrum of short-texts mining approaches and applications. Hence, this paper provides a comprehensive survey of this spectrum based on a criterion-based research strategy. The different mining techniques and approaches utilized in short-texts were highlighted along with their related issues and challenges. This paper surveyed a total of 1575 research papers published in the refereed conferences and journals in the area of short-texts mining were sur-veyed from 2006 until 2017, from which 187 primary studies were included and analyzed to constitute the source of the present paper. After a careful review of these articles, it is obvious that there are research gaps in other languages than English and Chinese, multi-languages, and in specific domain studies.

Keywords: Natural Language Processing, Arabic Language, Short Text, State of Art, Short Text Applications, Short Text Similarity

Introduction

The advent in the online space platforms and the popularity of micro-blogging, social networking and e-commerce systems, have attracted the attention to detect the needs, the tendencies and the opinions of and billions of online users expressed through short-texts such as messages, tweets and commodity reviews. Due to the paramount volume of short-texts produced daily, it is unfeasible to efficiently identify hidden knowledge patterns from such a massive mass of messages using traditional techniques. Therefore, it is necessary to develop automated methods to understand, summarize, classify and present all information in a clear and concise way. Due to the short-texts characteristics as sparsity, syntactical structure, noise and colloquial terminologies, the conventional machine learning and traditional text mining algorithms might not be the most appropriate tools for managing and analyzing large corpora of short-texts as it may not preserve the semantic meaning of the original texts. Therefore, the short-text mining techniques have become a focus research topic in recent years. Consequently, the literature was surveyed to present a systematic overview of the primary relevant studies on short text mining and to highlight the research gaps in the existing literature. Identifying such gap may guide the future research in such hot area toward the relevant under investigated areas. Moreover, providing a relevant clustering of the existing research is needed in such multi-discipline domain.

The rest of this paper is organized into eight sections as follows: After the sections of previous studies and the research methodology, sections four describes the sources of short text and section five addresses the preprocessing the representations techniques. Then the mining strategies and the intended application are introduced in sections six and seven before the survey is concluded in section eight.

Pervious Studies

Several studies addressed the challenges associated with short text modeling. Song *et al*. (2014) presented a review study summarizing the main characteristic of short texts, the main challenges facing short text classification and existing short text classification methods. However, they focused on classification and did not address any other mining strategy. Others review studies addressed only specific short-text applications. For example, Atefeh and Khreich (2015) presented another survey summarizing different event detection approaches for Twitter stream data. The approaches were classified according to the event type, the detection method and the detection task. In addition, they considered the commonly used feature representation and the target applications. In the same year, Injadat *et al.* (2015) presented another review of 66 articles derived with a criterion-based research strategy between 2003 and 2015 and concluded that there were 19 techniques for mining social media. Huang *et al.* (2016) conducted a comprehensive literature review based on case tracking and case studies for summarizing different research methods and challenging issues related to short text processing on Twitter. They concluded that text classification and text clustering are the key short text processing methods in microblogs, which are more applicable in many significant applications, including topic detection, sentiment analysis, smart healthcare, business intelligence and location services. They reported that semantic processing, noise reduction and handling big data are the main challenge issues for processing short texts in microblogs. To the best of our knowledge, there has not been any comprehensive survey addressing such critical and challenging area of research since 2016 despite the enormous number of short text studies published since that time. Besides, three of the studies mentioned above addressed the studies in social media, especially Twitter and microblogs, social media. The fourth study discussed only the short-text classification literature. Therefore, there is a need for a survey to cover the different aspects of short-text processing models and applications other than Twitter and the recent progress in this area since. The literature survey introduced by this research article attempts to find an answer to significant research questions, including:

- Which proper anatomy is needed to cluster the short text literature?
- What types of short-texts were considered in the literature (language, source)
- Which approaches are suitable for representing short texts
- Which strategies are commonly used for mining short texts
- What are the main applications domains of short text mining

Research Methodology

The main phases of a typical short text processing model are illustrated in Fig. 1. Models usually start with text Corpus collection from various sources, followed by the pre-processing phase required for cleaning, preparing and representing the data in a suitable form for the mining phase. In the mining phase, the proper text mining strategy and the proper evaluation technique are selected to serve the final application phase.

To find out the studies that address the short-text mining area, the academic databases of Springer, Scopus, Science Direct, CiteSeerX library, Google Scholar, ACM Digital library and IEEE-Xplore were searched for relevant publications published from 2006 until 2017. The keyword of “short text” was used in combination with “mining,” “clustering,” “classification,” “feature extraction,” “representation,” or “similarity.” The initial search process resulted in 1575 studies. As shown in Fig. 2, the 1575 studies were screened to filter out the replicated studies and the non-journal and non-conferences articles to end up with 836 studies. The abstracts of this set were screened to identify their relevance. The full-text of the obtained 476 studies were checked to identify the suitability for further analysis. A total of 289 articles were considered irrelevant to the short-text mining area and were excluded to yield to a final set of 187 paper.

As shown in Fig. 3, only 6% of the studies (11 papers) were published during the first three years of the surveyed period; on the other hand, 45% of the studies (84 papers) were published during the late three years of the surveyed period. The selected papers are classified according to the above four main phases of short text processing:

- Data set collection
- Text pre-processing and representation approaches
- Mining strategies
- Applications

The Dataset

The first phase of a short text processing model is to determine the domain of the model or the dataset. As shown in Fig. 4, it is evident that the most significant percentage of the research papers focused on utilizing free available benchmarking datasets for their methodology improvements. While, other researchers focused on building their own datasets by collecting data from websites, news articles and social media, especially when there is no free available benchmarking dataset particularly for some languages such as Arabic.

As shown in Fig. 5, the English language is the language of most of the studies followed by the Chinese language. Sixteen papers (9% of the studies) used a multi-language dataset. Considering the different nature and structure of each language, it strongly suggested that addressing other languages than English and Chinese and providing techniques to handle multi-languages short-text are under-investigated research areas.

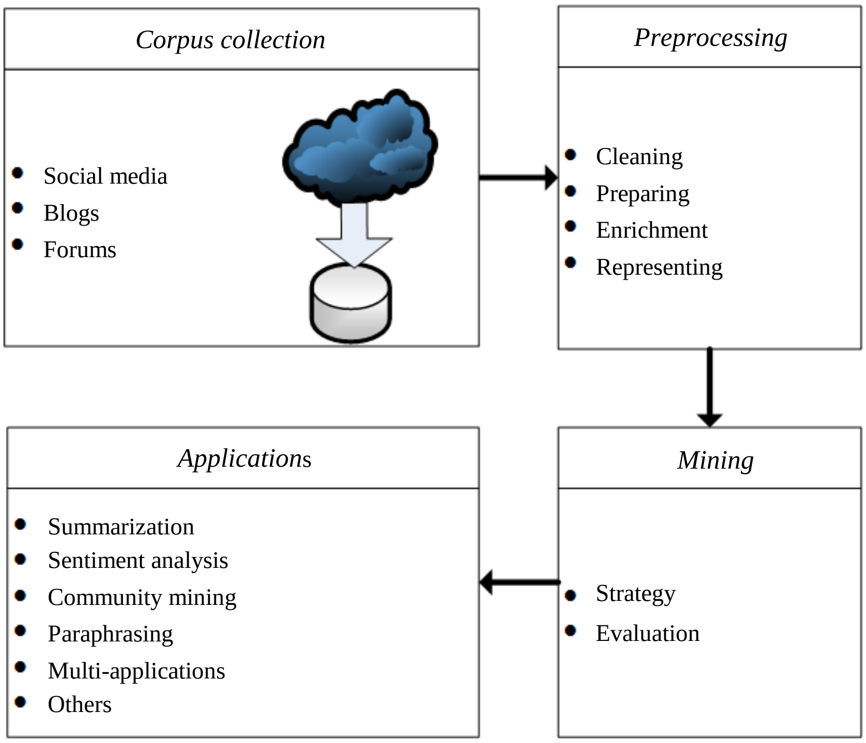

Fig. 1: Short-text processing phases

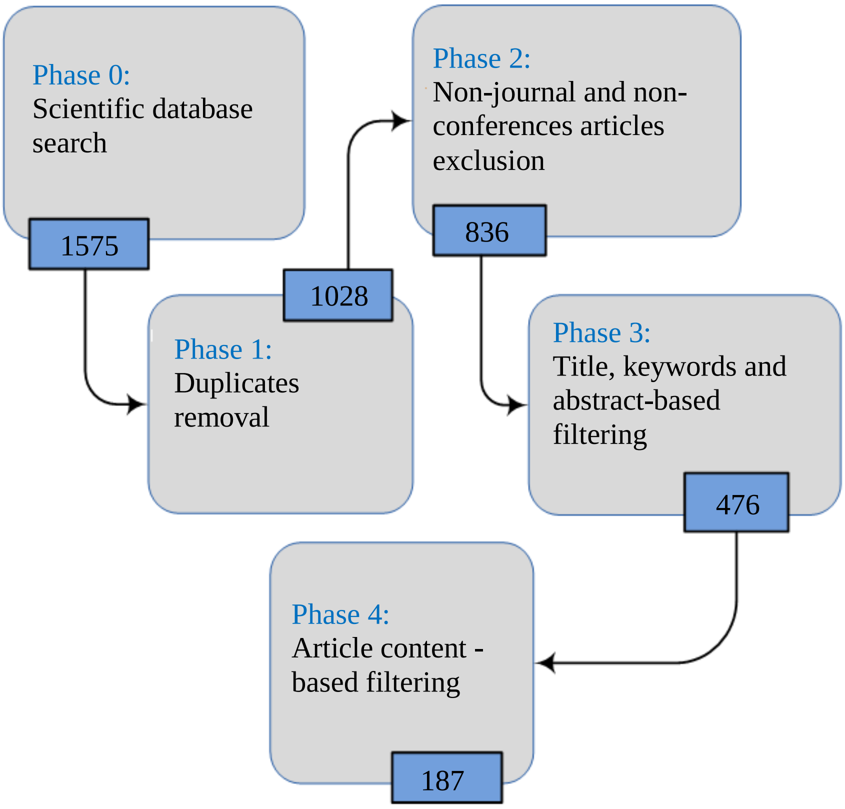
Phase 0:
Scientific database search
1575
Phase 1:
Duplicates removal
1028
Phase 2:
Non-journal and non-conferences articles exclusion
836
Phase 3:
Title, keywords and abstract-based filtering
476
Phase 4:
Article content - based filtering
187

Fig. 2: The number of studies in each phase of the selection and filtration process

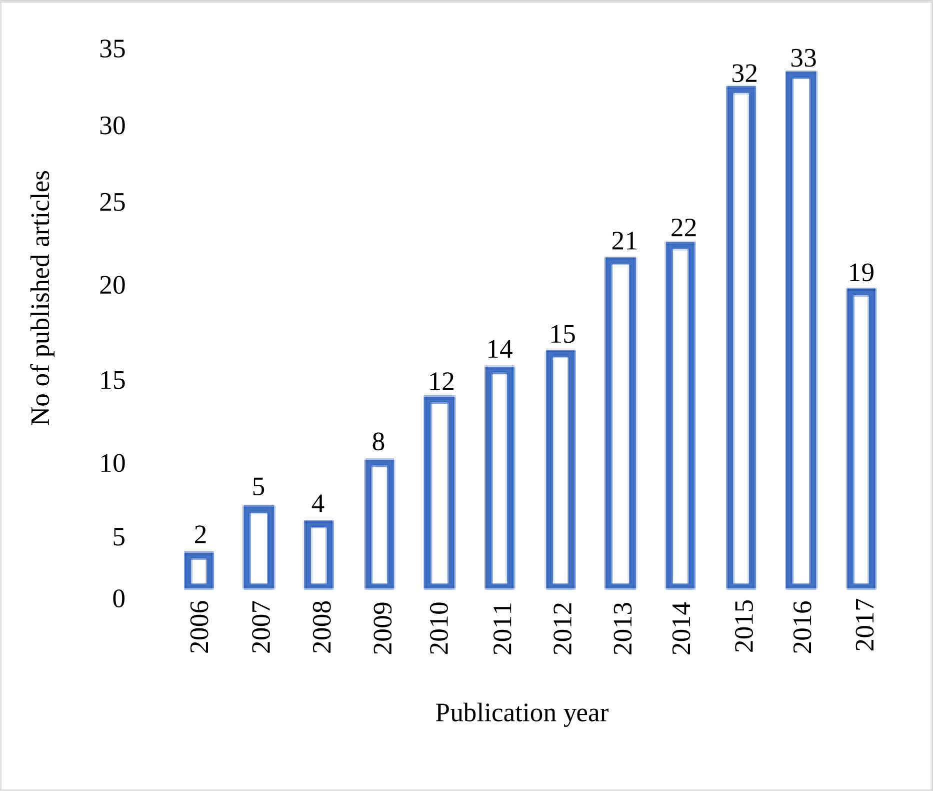

Fig. 3: The number of studies published in each year

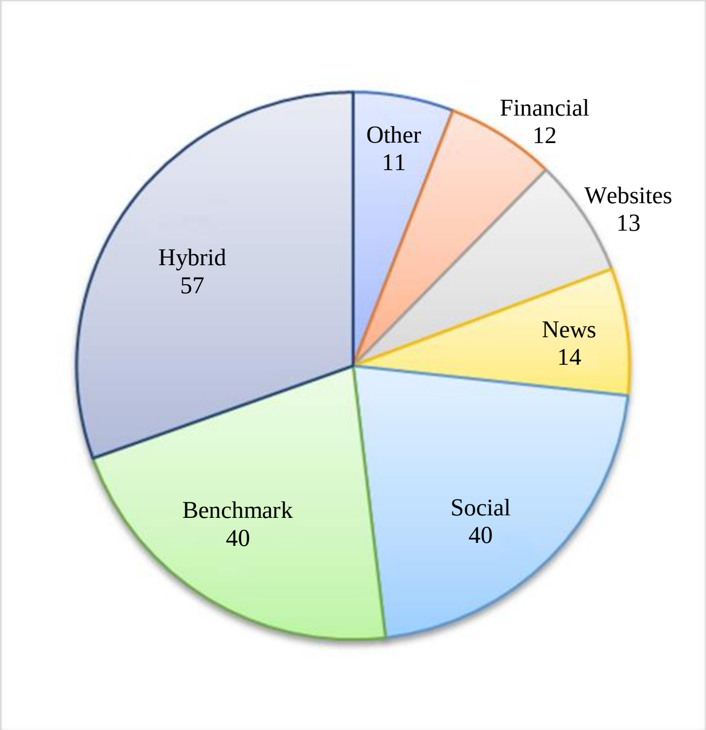

Fig. 4: The number of publications of each type of the data source

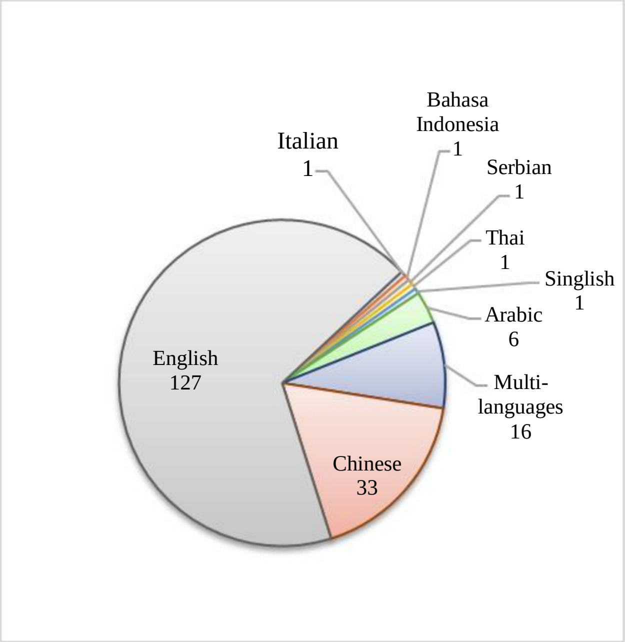

Fig. 5: The number of publications of each language of the data source

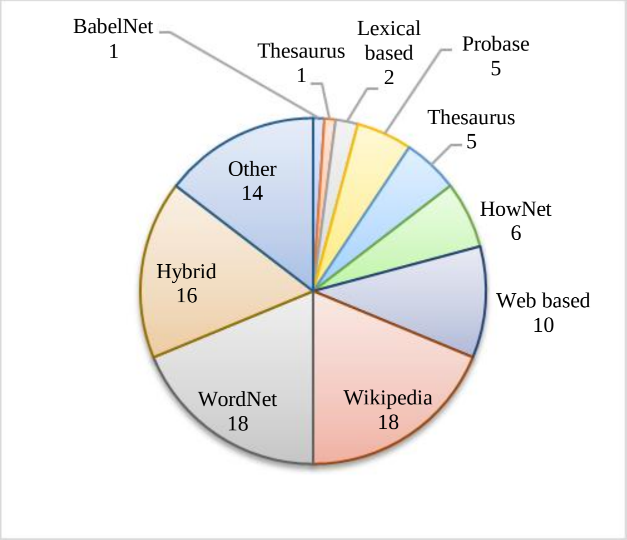

Fig. 6: The number of publications for each external knowledge resource

Preprocessing and Representation

This phase consists of four main steps as shown in Fig. 1. Despite that both the cleaning and preparation steps con-ducted for other data than short-text, the enrichment and the representation are essential only for short-text. Using the words cleaned short-text to represent it would underrepre-sent the data and would limit the mining results. Therefore, finding the appropriate enrichment technique along with the suitable representation to this technique is a corner stone in any short-text processing application.

The literature survey illustrated that pre-processing phase plays a critical role in the success of any short text mining technique as it prepares data and transforms it into a suitable format for pattern extraction and knowledge discovery. Preprocessing depends on several methods as tokenization for splitting text into words, normalization for removing any duplicate characters, stop word removal and lemmatization and stemming for returning the root of each word.

After pre-processing, the representation is conducted to produce the proper input of the analysis. The literature illustrated that the most popular text representation schema is Vector Space Model (VSM), which represents text with Bag of Words (BOW) vector ignoring semantic, syntactical, or lexical relations between input features. However, VSM is unsuitable for representing short text due to its noisy and sparsity problems.

Short text enrichment approaches were proposed to alleviate such problems by elevating the representation with either hidden statistical information using non-linguistic approaches or relevant hidden topics derived externally from large external sources such as search engines and Corpus. The short text enrichment methodologies can be classified into two main categories: linguistic and non-linguistic.

Linguistic Enrichment

Figure 6 shows a single external enrichment source; such as WordNet, Wikipedia, web search engines, HowNet, thesaurus, or Probase; are the most common sources for short text enrichment. It is evident that both WordNet and Wikipedia are the most used sources for short text enchainment. For instance, Abdalgader (2017) presented a synonym expansion semantic approach for enriching using additional semantic information aggregated from Word-Net lexical database; however, WordNet is not available for some languages. Nakamura *et al*. (2014) proposed two novel methods based on Wikipedia and Extended Naive Bayes (ENB) to enhance multilingual short texts clustering by incorporating inter-language links into ENB for unifying language and expanding short text representation with semantic information via a vector of Wikipedia entities.

Other studies enriched short-text with semantic features derived externally from a large-scale probabilistic knowledge base such as Probase (Hua *et al*., 2017). Those knowledge bases are used to identify semantic relatedness between terms, to segment the text, to detect its type and to generate concept labeling. Song *et al*. (2011) proposed a short text conceptualization method for enhancing short-text representation with additional fine-grained concepts derived from Probase and reflecting the most appropriate sense for each term under different contexts. Yu *et al*. (2016) presented an approach for short text understanding by combining semantic text enrichment and hashing. Their approach began with enriching each term of the short text with its relevant concepts and co-occurring terms inferred from Probase. After that, semantic short text hashing was performed through a deep neural network constructed based on a 3-layer stacked auto-encoders designed with a specific learning strategy. Alternatively, other researchers focused on topic analysis models to resolve the sparsity problem in short text. Chen *et al*. (2011) presented an approach for enriching short text representation and classification by mapping it into a feature space of multi-granularity topics derived from a large external corpus.

Non-Linguistic Enrichment

The non-linguistic literature tackles content sparsity problem of short-text modelling. Quan *et al*. (2015) introduced a Self-Aggregation-based Topic Model (SATM) by integrating topic modelling with clustering for text self-aggregation during topic inference, while others focused on word co-occurrence information as self-contained knowledge for enhancing topic modelling for sparse and short text. Chen and Kao (2017) attempted to solve the problem of inadequate word co-occurrence patterns in a short document through a Re-Organized document LDA (RO-LDA). The RO-LDA is word co-occurrence improvement method, which can extend the length of each short text document by re-organizing its words into a virtual document using the word co-occurrence information in the whole corpus.

Mining Strategy

Setting the strategy of short-text mining requires to decide between classification and clustering, select the appropriate model and to select the proper similarity method.

Classification and Clustering

Text classification is defined as a technique for assigning the text into a number of predefined categories. However, due to the sparsity and noise of the short text, the traditional text processing models encounter are not satisfactory effective with short text classification. On the other hand, text clustering is defined as a technique for grouping similar text documents in the same class using similarity measure. The literature illustrated that the quality of text-clustering methods relies on three aspects, including text representation model, similarity metric and clustering algorithm applied to group texts. Existing models of short text classification and clustering are mainly based on the prior text enrichment technique.

Linguistic Based Models

Some short text classification studies enriched the text feature space with topics and knowledge derived from estimated topic models using Latent Dirichlet Allocation (LDA) but from three types of universal large-scale corpora: Wikipedia, DBLP and LNCS (Vo and Ock, 2015). Other studies succeeded in improving short text classification by encoding semantic information of short text via word embedding based approach, which pre-trained over large-scale external Corpus. For instance, Wang *et al*. (2016) proposed a convolutional short text modeling for enhancing short text classification by expanding short text features with the multi-scale semantic information discovered from pre-trained words embeddings and Convolutional Neural Network (CNN). Similarly, clustering studies relied on auxiliary data from a large-scale external corpus, such as Wikipedia or web search engine to overcome the sparsely problem in short text clustering (Hu *et al*., 2017; Zhou *et al*., 2016). Despite the improvement in short text classification and clustering obtained by the above models, they have some limitations. Externally enriched models are time-consuming and heavily dependent on the quality of the external source (search engine, large-scale Corpus, etc.) Moreover, such models may face a significant challenge to find strongly related external corpora to a specific short text of a particular domain, such as military due to privacy or confidentiality reasons. On the other side, knowledge enriched models are dependent on semantic features derived from external taxonomy or lexical base, which are often unavailable for many languages.

Non-Linguistic Based Models

Due to previously described limitations of linguistic methods, some studies utilized latent semantic analysis to enhance performance of short text classification by en-richening short text with self-contained knowledge. They extracted such knowledge directly from the text Corpus itself either via clustering (Dai *et al*., 2013) or statistical co-occurrence analysis between terms (Rao *et al*., 2016). The literature showed that feature extension approaches, which are based on term co-occurrence, may be ineffective especially when terms from different sentences have a weak semantic relation. Kim *et al*. (2014) short text classification via a Language-Independent Semantic kernel (LIS) based on three levels of semantic annotations for extending short text with semantic and syntactic features without using ready-made lexical databases and grammatical tags. Such approaches can only achieve satisfactory results with a broad training Corpus to extract meaningful associations. Consequently, other approaches for enhancing short text classification were presented when the training dataset is small or fewer word co-occurrences information among terms are found. For instance, Ramírez-De-La-Rosa *et al*. (2013) proposed a Neighborhood-Consensus Categorization (NCC) method for classifying short text documents to determine the category of each document by considering its content and information about the class of the neighborhood documents.

Similarity Estimation

Finding how similar a short message to another one is a cornerstone of the mining strategies. Similarity measures can be categorized to linguistically based measures and non-linguistic based ones. The simple cosine of the angle between two arrays representing two short texts is the most common lexical similarity measure. Euclidean distance, Jaccard similarity and Pearson's correlation coefficients are commonly used measures in the literature as well.

Lee (2011) proposed a sentence similarity measure based on extracted tags as joint noun set and joint verb set and represented as WordNet's semantic tree. Alzahrani (2016) estimated the similarity between cross-language short texts through three proposed semantic similarity algorithms: Averaged Maximum-Translation Similarity, Noun-Verb Vector Based Similarity and Machine Translation (MT) Term Vector Based Similarity using WordNet semantic information. Generally, the knowledge-based methods are limited to the human-crafted dictionaries, which do not include all words and do not have the required semantic information of the included words, e.g., WordNet has a fewer number of verbs and adverbs synsets than the nouns synsets.

Batanovic and Bojic, (2015) proposed a short text semantic similarity method, named POST STSS method. The method based on the combination of a statistical bag-of-words approach and part-of-speech tags as more profound syntactic information indicators without any hand-crafted knowledge base or advanced syntactic tools. Liu *et al*. (2016) introduced a model for assessing short text similarity based on double vector space model constructed using semantic information from Wikipedia in addition to structure information from a semantic dependency tree.

Furlan *et al*. (2013) presented a statistical language independent-based approach for estimating the semantic similarity between short texts. This approach was based on statistical distribution of words in short text corpus itself for evaluating semantic short text similarity by combining word-to-word string and semantic similarities in addition to utilizing a weighted term frequency ponderation for enhancing final similarity score.

Applications of Short Text Mining

Short text mining was used for different applications such as summarization, topic detection, event detection, information retrieval, sentiment analysis, community mining and recommendation systems, etc. as shown in Fig. 7. Despite that the many of the surveyed papers addressed only the short-text mining methodology improvement, one hundred and sixty-five of them focused on a specific application area.

Paraphrase Identification

The Paraphrase identification is as a classic natural language processing task, which takes a pair of sentences as an input to check whether they are paraphrase of each other by estimating semantic relatedness among them.

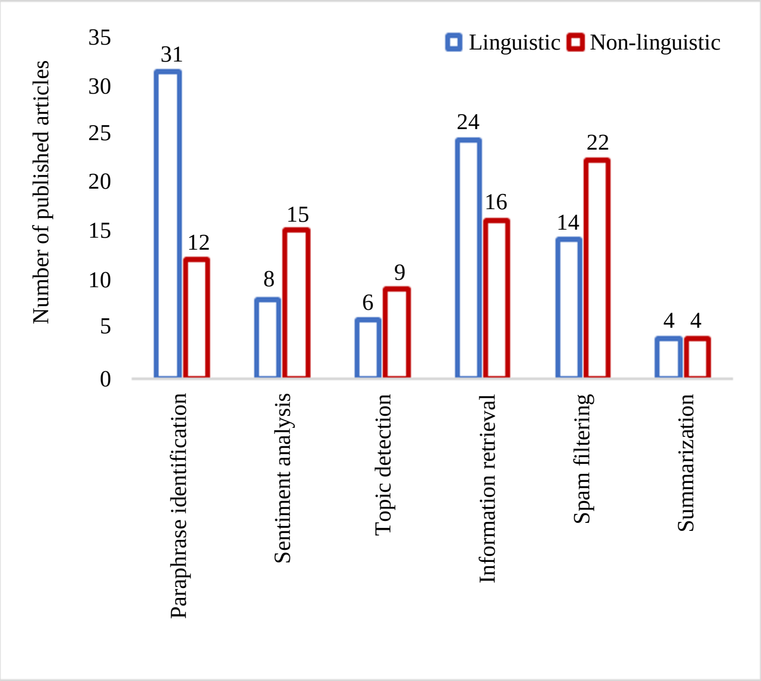

Fig. 7: Number of published articles in each application area

The literature illustrated that paraphrase identification is a base for other short text mining application such as summarization, word sense disambiguation, plagiarist, information retrieval and machine translation. Researchers of paraphrase identification focused more on linguistic approaches for processing short texts (31 papers) more than non-linguistic processing approaches (12 papers).

Sentiment Analysis

Recently, sentiment analysis or opinion mining has become an essential research area in a variety of domains, including commerce, education, health, tourism and politics.

Tripathy *et al*. (2016) utilized the n-gram machine learning approach to classify human sentiments in movie reviews. Sun *et al*. (2016) presented a method for improving sentiment classification of Chinese microblog posts by combing each post with its comments and then extract sentiment related information via a proposed deep neural network model with a content extension method. Rao *et al*. (2016) presented a Topic-Level Maximum Entropy (TME) model for classifying social emotions by combining all word tokens with latent topics extracted from short text via unsupervised topic models by modeling the word co-occurrence patterns to alleviate problems in the ME principle. Recently, alternative few approaches for short text opinion mining have been introduced based on primarily linguistic approaches. For instance, Lochter *et al*. (2016) presented an ensemble system for enhancing opinion detection over social generated short texts. They integrated some text processing approaches (including text normalization and semantic indexing for short normalization and expansion) with traditional classification methods for polarity detection over short text messages.

Topic Detection

With the popularity of various Online Social Networks (OSNs), they become the primary resources for spreading valuable information in the context of education and entertainment in addition to political campaigning and news reporting. However, they still have a negative side as they can disseminate unverifiable information and rumors. Therefore, the literature focused on microblog topic detection for analyzing massive short text streams and detecting rumors by identifying memes (e.g., a unit of information spread among users through OSNs). The majority of the reviewed papers in topic detection (twelve out of fifteen) were based on unsupervised learning methods (clustering) with 12 reviewed papers.

Traditionally, the literature focused on topics mining from a single language social media data (mostly English); recently, some researchers presented a multilingual approach for online social media topic identification. For instance, Lo *et al*. (2017) introduced a multilingual approach for detecting extremely relevant terms and essential topics from the enormous social media data by integrating localized language analysis and proposed 'Joint' term ranking method with unsupervised topic clustering and multilingual sentiment analysis for topics ex-traction via analysis of Twitter's tweets during a period of time.

Summarization

With the rapid advance of social networks, a massive amount of social contents is generated continuously reflecting all online users' feelings or opinions toward a variety of entities, including products, services, topics, persons and organizations.

Hu *et al.* (2017) presented an approach for summarizing online hotel reviews via sentence importance metric for identifying most informative sentence from reviews based on not only their contents information but also on other critical contained factors, including authors credibility, time, review usefulness and conflicting opinions. They used the k -medoids clustering algorithm to partition important sentences into k-clusters after estimating both content and sentiment similarities among them. Then the top-k sentences were selected as the final summarization result after estimating important sentence score for each cluster. Instead of exploiting a clustering-based approach for obtaining short text summarization Amplayo and Song, (2017) relied on constructing multi-level sentiment classification model and aspect extraction model for presenting a fine-grained sentiment extraction model to enhance the summarization of multiple online reviews.

Information Retrieval

Information retrieval is one of the critical research areas in a variety of applications, including E-commercial, E-government, news and academic applications, which can help users to search and gain their required contents. Yu *et al*. (2016) attempted to enhance Information Retrieval (IR) task on the MSN news data via an approach for short text understanding by combining semantic text enrichment and hashing. They began by enriching each term in input short text with its relevant concepts and co-occurring terms inferred from Probase. After that, semantic short text hashing was performed through a deep neural network constructed based on a 3-layer stacked auto-encoders designed with a specific learning strategy.

Spam Filtering

With the rapid evolution of short text messages in recent electronic communication media, a variety of undesired contents (spam) can be propagated among users, including misleading information, ads, viruses, which may be harmful and need to be detected and filtered. Silva *et al.* (2017) introduced an incremental short text classification learning approach based on the Minimum Description Length (MDLText) for online spam detection and filtering over short and sparse short text messages.

Conclusion

With the rapid and massive amount of unstructured short text daily feed, the ability to extract valuable information from these data becomes very beneficial for organizations and individuals for different and essential purposes, including opinion mining, web search result categorization, event detection, text summarization, spam detection, etc. As a result of systematical literature survey for the between the period 2006 and 2017, 187 relevant research papers were identified. Despite the progress of the short texts mining applications, the analysis of, it was concluded that there is still some of research gaps and opportunities in the existing researches which are detailed as follows:

- There is a lack of studies dealing with informal texts mostly used among the youth on the internet and with languages other than English and Chinese
- There is a significant challenge to find strongly related external corpora to specific short text in a particular domain, such as military, healthcare and industrial applications
- There is a significant shortage in the research handling short text in multi-languages, despite the popularity of mixing English or French with other languages

Moreover, the paper introduced a structured clustering scheme of the existing research based on the source of the short-text dataset, the test representation and the enrichment methodology, the mining strategy, and the application domain. In general, it is research worthy to consider a detailed study to compare two or more of the introduced short-text mining approaches and short-text separation algorithms.

Acknowledgement

The author(s) received no specific funding for this work.

Author's Contributions

Mohamed Grida: Designed the paper classification scheme. Conduced the analysis, wrote the paper.

Hasnaa Soliman: Collected the reviewed papers, filtered them, and filled the paper classifications table and comments.

Mohamed Hassan: Revised the manuscript, designed the research plan.

Ethics

The authors declare that there is no conflict of interests regarding the publication of this article/ paper.

References

Abdalgader, K., 2017. Clustering short text using a centroid-based lexical clustering algorithm. IAENG Int. J. Comput. Sci., 44: 523-536. DOI: 10.5772/intechopen.75433

Alzahrani, S., 2016. Cross-language semantic similarity of Arabic-English short phrases and sentences. J. Comput. Sci., 12: 1-18. DOI: 10.3844/jcssp.2016.1.18

Amplayo, R.K. and M. Song, 2017. An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews. Data Knowl. Eng., 110: 54-67. DOI: 10.1016/j.datak.2017.03.009

Atefeh, F. and W. Khreich, 2015. A survey of techniques for event detection in Twitter. Computat. Intelli., 31: 133-164. DOI: 10.1111/coin.12017

Batanovic, V. and D. Bojic, 2015. Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity. Comput. Sci. Inform. Syst., 12: 1-31. DOI: 10.2298/CSIS131127082B

Chen, G.B. and H.Y. Kao, 2017. Word co-occurrence augmented topic model in short text. Intell. Data Anal., 21: S55-S70. DOI: 10.3233/IDA-170872

Chen, M., X. Jin and D. Shen, 2011. Short text classification improved by learning multi-granularity topics. Proceedings of the 22th International Joint Conference Artificial Intelligence, Jul. 16-22, Spain, pp: 1776-1781.

Dai, Z., A. Sun and X.Y. Liu, 2013. CREST: Cluster-based representation enrichment for short text classification. Proceedings of the Pacific-Asia Conference Knowledge Discovery Data Mining, (DDM. 13), pp: 256-267. DOI: 10.1007/978-3-642-37456-2_22

Furlan, B., V. Batanović and B. Nikolić, 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. Dec. Support Syst., 55: 710-719. DOI: 10.1016/j.dss.2013.02.002

Hu, Y.H., Y.L. Chen and H.L. Chou, 2017. Opinion mining from online hotel reviews-a text summarization approach. Inform. Process. Manage., 53: 436-449. DOI: 10.1016/j.ipm.2016.12.002

Hua, W., Z. Wang, H. Wang, K. Zheng and X. Zhou, 2017. Understand short texts by harvesting and analyzing semantic knowledge. IEEE Trans. Knowl. Data Eng., 29: 499-512. DOI: 10.1109/TKDE.2016.2571687

Huang, W., Z. Li, L. Zhang and Y. Li, 2016. Review of intelligent microblog short text processing. Web Intell., 14: 211-228. DOI: 10.3233/WEB-160340

Injadat, M., F. Salo and A. Bou, 2016. Data mining techniques in social media: A survey. Neurocomputing, 214: 654-670. DOI: 10.1016/j.neucom.2016.06.045

Kim, K., B. Chung, Y. Choi, S. Lee and J.Y. Jung *et al.*, 2014. Language independent semantic kernels for short-text classification. Expert Syst. Applic., 41: 735-743. DOI: 10.1016/j.eswa.2013.07.097

Lee, M.C., 2011. A novel sentence similarity measure for semantic-based expert systems. Expert Syst. Applic., 38: 6392-6399. DOI: 10.1016/j.eswa.2010.10.043

Liu, Y., D. Li and C. Dai, 2016. Short text similarity measure based on double vector space model. Int. J. Database Theory Applic., 9: 33-46. DOI: 10.14257/ijdta.2016.9.10.04

Lo, S.L., R. Chiong and D. Cornforth, 2017. An unsupervised multilingual approach for online social media topic identification. Expert Syst. Applic., 81: 282-298. DOI: 10.1016/j.eswa.2017.03.029

Lochter, J.V., R.F. Zanetti, D. Reller and T.A. Almeida, 2016. Short text opinion detection using ensemble of classifiers and semantic indexing. Expert Syst. Applic., 62: 243-249. DOI: 10.1016/j.eswa.2016.06.025

Nakamura, T., M. Shirakawa, T. Hara and S. Nishio, 2014. Semantic similarity measurements for multi-lingual short texts using Wikipedia. Proceedings of the International Joint Conference Web Intelligence Intelligent Agent Technology - Workshops, Aug. 11-14, IEEE Xplore Press, Warsaw, Poland, pp: 22-29. DOI: 10.1109/WI-IAT.2014.76

Quan, X., C. Kit, Y. Ge and S.J. Pan, 2015. Short and sparse text topic modeling via self-aggregation. Proceedings of the 24th International Conference Artificial Intelligence, Jul. 25-3, Buenos Aires, Argentina, pp: 2270-2276.

Ramírez-de-la-Rosa, G., M. Montes-y-Gómez, T. Solorio and L. Villaseñor-Pineda, 2013. A document is known by the company it keeps: Neighborhood consensus for short text categorization. Lang. Res. Evaluat., 47: 127-149. DOI: 10.1007/s10579-012-9192-1

Rao, Y., H. Xie, J. Li, F. Jin and F.L. Wang *et al.*, 2016. Social emotion classification of short text via topic-level maximum entropy model. Inform. Manage., 53: 978-986. DOI: 10.1016/j.im.2016.04.005

Silva, R.M. T.C. Alberto, T.A. Almeida and A. Yamakami, 2017. Towards filtering undesired short text messages using an online learning approach with semantic indexing. Expert Syst. Applic., 83: 1-41. DOI: 10.1016/j.eswa.2017.04.055

Song, G., Y. Ye, X. Du, X. Huang and S. Bie, 2014. Short Text Classification: A Survey. J. Multimedia, 9: 635-643. DOI:10.4304/jmm.9.5.635-643

Song, Y., H. Wang, Z. Wang, H. Li and W. Chen, 2011. Short text conceptualization using a probabilistic knowledgebase. Proceedings of the 30th International Joint Conference Artificial Intelligence, Jul. 16-22, Barcelona, Catalonia, Spain, pp: 2330-2336.

Sun, X., C. Li and F. Ren, 2016. Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features. Neurocomputing, 210: 227-236. DOI: 10.1016/j.neucom.2016.02.077

Tripathy, A., A. Agrawal and S.K. Rath, 2016. Classification of sentiment reviews using n-gram machine learning approach. Expert Syst. Applic., 57: 117-126. DOI: 10.1016/j.eswa.2016.03.028

Vo, D.T. and C.Y. Ock, 2015. Learning to classify short text from scientific documents using topic models with various types of knowledge. Expert Syst. Applic., 42: 1684-1698. DOI: 10.1016/j.eswa.2014.09.031

Wang, P., B. Xu, J. Xu, G. Tian and C.L. Liu *et al*., 2016. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174: 806-814. DOI: 10.1016/j.neucom.2015.09.096

Yu, Z., H. Wang, X. Lin and M. Wang, 2016. Understanding short texts through semantic enrichment and hashing. IEEE Trans. Knowl. Data Eng., 28: 566-579. DOI: 10.1109/TKDE.2015.2485224

Zhou, X., X. Wan and J. Xiao, 2016. CMiner: Opinion extraction and summarization for Chinese microblogs. IEEE Trans. Knowl. Data Eng., 28: 1650-1663. DOI: 10.1109/TKDE.2016.2541148