RESEARCH ARTICLE OPEN ACCESS

Capitalising on Wikipedia's Big Data for Text Analytics

Mohamed Minhaj

Research Scholar, International School of Information Management, University of Mysore - Mysore

ABSTRACT

While a lot of data, predominantly unstructured and textual in nature, is available in most organisations, they are not able to harvest and harness actionable information from that data. Text analytics deals with deriving novel, relevant and interesting patterns from data stored in organisations in the form of reports, weblogs, emails, social media etc. Typical text analytics tasks include categorisation, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarisation etc. Most of these tasks require access to an extensive collection of updated and quality information either to handle the ambiguity associated with the natural language text or to train and test the Machine Learning models related to text analysis. In this context, Wikipedia has a great potential to be used as background knowledge for natural language processing and other text analytics applications.This paper endeavours to explore the possible ways of leveraging the vast, updated and collaboratively created knowledge of Wikipedia for different text analytics tasks.

I. INTRODUCTION

Text analytics has gained a great deal of attention in recent years due to the tremendous amount of text data generated each day in various forms and the business value that the organisations have identified in such textual sources[1]. The text data available in the form of reports, news archives, scientific articles, blogs, emails, social networks etc., is a treasure trove of business insights. However, the volume, heterogeneity and unstructured nature of text data make its consumption a challenging task. This phenomenon has fueled the research on Text Analytics. Text analytics is the process of deriving high-quality, actionable information/insights from text. Text analytics generally involves structuring the input text, deriving patterns within the structured data, and finally, evaluating and interpreting the output. 'High quality" in text analytics refers to some combination of relevance, novelty, and interestingness. Typical text analytics tasks include text categorisation, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarisation etc.

Text analytics comprises a collection of machine learning, linguistic and statistical techniques used to model and extract information from text primarily for analysis needs, including business intelligence, exploratory, descriptive and predictive analytics[2]. Many text analytics tasks require access to large amounts of secondary or background data to understand the text being analysed or train machine learning models. Among the plethora of textual sources available on the web, Wikipedia is one of the prominent sources and can play a pivotal role in Text Analytics. Wikipedia has 51 million articles, with 6 million articles in English alone. However, users are not able to use the full potential of Wikipedia's big data because of constraints like limited full-text search and machine interpretation. Hence, besides studying the direct use of Wikipedia's knowledge base, there are several research opportunities to use Wikipedia to support text analytics activities.

II. RESEARCH OBJECTIVES AND METHODOLOGY

Today, most organisations are in a "data-rich and information poor" state. While a lot of data, predominantly unstructured and textual in nature, is available in most organisations, they are not able to harvest and harness actionable information from that data. Text analytics deals with extracting novel, relevant and interesting patterns from data stored in organisations in the form of reports, weblogs, emails, social media etc. Many text analytics tasks, such as Sentiment Analysis, Text Classification, Summarisation, etc., require access to an extensive collection of quality background information to handle the ambiguity associated with the natural language text or for training and testing the Machine Learning models.

This paper endeavours to study the possible ways of capitalising on the vast, updated and collaboratively created knowledge of Wikipedia for different text analytics tasks. The study is exploratory in nature and is based on the recent literature ferreted from scholarly databases. The intended outcome of this study is to critically appraise and summarise the existing evidence concerning the use of Wikipedia for Text Analytics and aid future research.

The remainder of this paper is organised as follows: Section 3 highlights the conceptual background of Text Analytics. Section 4 presents the significance of Wikipedia for Text Analytics. Section 5 describes how Wikipedia is being capitalised for different Text Analytics tasks, and Section 6 concludes the paper.

III. CONCEPTUAL BACKGROUND OF TEXT ANALYTICS

In the present milieu of business, organisations are generating data at incredible speed and volume, much of which is

unstructured. While organisations are using structured data stored in their databases and spreadsheets for their day-to-day requirements, they find gaining insights from unstructured sources challenging. As the conventional analytical techniques and statistical methods used to get insights from structured data are not suitable for unstructured text data, Text Analytics has gained prominence.
Text analytics, also known as text mining or knowledge discovery from text, is the methodology and process to derive quality and actionable information and insights from textual data. Text mining was first introduced by Fledman et al.[3]. It involves using Natural Language Processing (NLP), Information Retrieval (IR) and Machine Learning (ML) techniques to parse unstructured text data into more structured forms and derive patterns and insights that would be helpful for the end-user. Text analytics entails a collection of techniques used to model and extract information from text primarily for analysis needs, including business intelligence, exploratory, descriptive and predictive analytics[2].

Text representation and encoding :
Text mining from large textual documents is a complex process, and thus it is critical to have a data structure for the text which facilitates further analysis of the documents [4]. The most common approach to represent the textual data is Bag of Words (BoW). A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves a vocabulary of known words/terms and a measure of the presence of known words/terms. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. This representation leads to a vector representation referred to as Term by Document Matrix (Shown in Table 1) and can be used to analyse the text using machine learning and statistics.

Table 1 Term by Document Matrix

	Term 1	Term2	Term 3	…	Term M
Document 1					
Document 2					
Document 3					
…					
Document N					

Text Preprocessing :
Pre-processing is one of the key tasks in Text Analytics and impacts the quality of any text analytics outcome. For example, a traditional text categorisation framework comprises preprocessing, feature extraction, feature selection and classification steps. While it is evident from several studies that feature extraction, feature selection, and classification algorithm have an impact on the classification process, the preprocessing stage has exhibited a significant influence on the success of text categorisation. Uysal et al. have investigated the impact of preprocessing tasks, particularly in text classification[5]. The preprocessing step usually consists of the tasks such as tokenisation, filtering, lemmatisation and stemming.

Tokenisation is the first step in text analytics and is the process of breaking down a text into smaller chunks such as words or sentences called tokens.
Filtering refers to the removal of certain words from the text being processed. The most common filtering is stop-word removal. Stop words appear frequently but are very generic in meaning and do not contain much information (Ex. prepositions, conjunctions, etc.).
Stemming is the process of eliminating affixes (suffixed, prefixes, infixes, circumfixes) from a word to obtain a word stem. Ex. running → run.
Lemmatisation is related to stemming and focuses on capturing canonical forms based on a word's lemma. It involves grouping together the various inflected forms of a word to be analysed as a single item. In other words, lemmatisation methods map verb forms to infinite tense and nouns to a single form.

Text Analytics Approaches:
Information Retrieval (IR): It is the activity of finding information resources (usually documents) from a collection of unstructured data sets that satisfies the information need[6]. Information Retrieval deals with facilitating information access rather than analysing information and finding hidden patterns, which is the primary purpose of mining.
Natural Language Processing (NLP): It is a field of study that combines computer science, artificial intelligence, and linguistics to understand the natural language using computers[7]. Many text mining algorithms extensively use NLP techniques, such as parts of speech tagging (POS), syntactic parsing, and other types of linguistic analysis.
Text Summarisation: Many text mining applications need to summarise the text documents to get a concise overview of a large document or a collection of documents on a topic [8]. There are two categories of summarisation techniques in general: extractive summarisation, where a summary comprises information units extracted from the original text, and contrary abstractive summarisation, where a summary may contain "synthesised" information that may not occur in the original document [9].
Information Extraction from the text (IE): Information Extraction is the task of automatically extracting information or facts from unstructured or semi-structured documents[10]. It usually serves as a starting point for other text mining algorithms[9]. For example, extracting entities and their relations from the text can give us useful semantic information.
Machine Learning (ML): A subset of AI provides computing systems with the ability to learn without being explicitly programmed. ML focuses on the development of Computer Programs that can change when exposed to new data. As far as analytics is concerned, ML is a method of data analysis that

automates analytical model building and works on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Based on the different approaches used to train the machines or how the system is modelled to discover the patterns from data, machine learning can be classified as Supervised and Unsupervised.

Supervised Learning: When an algorithm learns from example data, and associated target responses that can consist of numeric values or string labels, such as classes or tags, to later predict the correct response when posed with new examples, such type of machine learning is referred to as Supervised[11].

Ex. Text Classification

Unsupervised Learning: The type of ML where an algorithm learns from simple examples without any associated response, leaving the algorithm to determine the data patterns on its own, is referred to as unsupervised learning.

Ex. Text Clustering, Topic Modeling

Text Classification :

Text classification, also known as text tagging or text categorisation, categorises text into organised groups. Using Natural Language Processing (NLP), text classifiers can automatically analyse text and then assign a set of pre-defined tags or categories based on its content [12]. Common examples of Text Classification include Sentiment Analysis, Topic Detection etc.

Text Clustering:

It is the application of cluster analysis to text-based documents. It uses machine learning and natural language processing (NLP) to understand and categorise unstructured textual data. Typically, descriptors (sets of words that describe topic matter) are extracted from the document first. Then they are analysed for the frequency in which they are found in the document compared to other terms. After which, clusters of descriptors can be identified and then auto-tagged.

IV. WIKIPEDIA: THE BIG DATA REPOSITORY

Wikipedia, the big data of text, facts and figures, is a multilingual, web-based, free-content encyclopaedia project supported by the Wikimedia Foundation. Wikipedia has been the most successful collaborative encyclopaedia, and its English edition alone has more than 6 million articles. The key facts about Wikipedia's richness are shown in Figure 1.

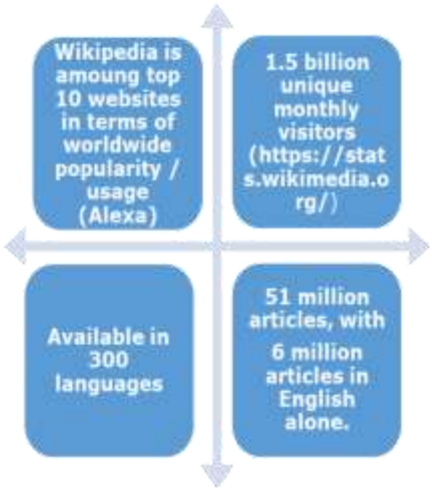
Wikipedia is amoung top 10 websites in terms of worldwide popularity / usage (Alexa)
1.5 billion unique monthly visitors (https://stats.wikimedia.org/)
Available in 300 languages
51 million articles, with 6 million articles in English alone.

Figure 1 Key facts about Wikipedia's richness

The Wikipedia contributors, in addition to the quantity, strive to improve the quality. Every article in Wikipedia is considered to be in a work-in-progress phase and progresses to various stages of completion. As articles develop, they tend to become more comprehensive and balanced. Quality also improves over time as misinformation and other errors are removed or rectified.

In addition to a large amount of unstructured data, Wikipedia articles also have structured data enveloped in them in the form of Infoboxes. An Infobox is a fixed-format table usually added to the top right-hand corner of articles. The infoboxes summarise essential points in an easy-to-read format and also improve navigation to other related articles. With the structured nature of data and the possibility of mapping its schema efficiently to many dominant metadata systems, the data in Wikipedia's infoboxes is widely used for many knowledge-based applications. The notable applications that are built primarily on Wikipedia's infoboxes include DBPedia.

V. WIKIPEDIA AS A RESOURCE FOR TEXT ANALYTICS

Text Analytics refers to the discipline of Computer Science, which employs Natural Language Processing and Machine Learning to draw meaning and insights from unstructured text data. The prominent Text analytics tasks include Information Retrieval, Information Extraction, Text Summarisation, Text Classification etc.

The data used by the organisations for the tasks mentioned above include emails, weblogs, reviews, feedback, survey data, data harvested from social networks such as Facebook, Twitter etc. However, many text analytics tasks require access to large amounts of secondary or background data either to understand the text being analysed or to train machine learning models. In such a context, a vast amount of collaboratively curated knowledge of Wikipedia can play a pivotal role.

In the light of recent scholarly literature, the works involving Wikipedia for any form of Text Analytics have been explored and presented in the following section. The different categories of text analytics tasks that have leveraged Wikipedia knowledge have been depicted in figure 2.

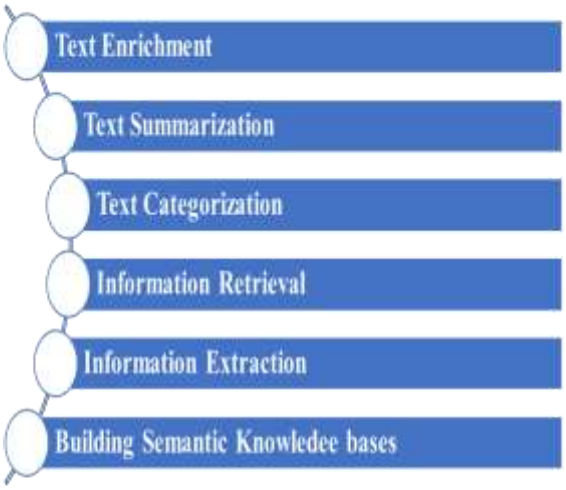

Figure 2 Different categories of text analytics tasks that have leveraged on Wikipedia's knowledge

A. Text Enrichment using Wikification

The natural language text available in the form of news items, journal articles, course reading materials, blog entries or social media posts is associated with a lot of ambiguity. While the degree of difficulty in reading and understanding such texts may vary depending on the reader, any form of background or secondary information about the text, made accessible contextually in a timely and non-intruding manner, can make a big difference in helping the readers in comprehending the text.

When people have any difficulty understanding a particular term/word in the text document, generally, the user may search the meaning/definition of the difficult term in a search engine like Google. Further, it has been observed that in most cases, the first page in the search engines result page is Wikipedia. Therefore, it would help the readers if the key terms in the text being read were automatically linked to the relevant pages on Wikipedia. This process is predominantly referred to in the scholarly literature as Wikification.

Given an input document, the Wikification system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages [13].

B. Text Summarisation

Text summarisation refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. Automatic text summarisation is an essential task in machine learning and natural language processing (NLP).

Several attempts have been made to leverage Wikipedia for text summarisation. In work by Ramanathan, Sankarasubramaniam, Mathur, and Gupta, the document sentences are mapped to semantic concepts in Wikipedia and the sentences are selected for a summary based on the frequency of the mapped concepts[14]. The work by Ye, Chua, and Lu is focused on summarising definitions using Wikipedia pages [15]. One of the other prominent works is by Pourvali & Abadeh, which leverage Wikipedia to form multiple independent graphs, and then use graph importance and lexical cohesion features for summarisation [16].

C. Text Categorisation

Text classification, also known as text categorisation or tagging, involves analysing a piece of text and assigning a set of pre-defined categories. Text classifiers can be used to organise, structure, and categorise any type of textual data. For example, news articles can be organised by topics, support tickets can be organised by urgency, chat conversations can be organised by language, brand reviews can be organised by sentiment, and so on.

For example, if a customer review about a product is having the following text:

"The product has excellent features and is value for money."

A classifier can take this text as an input, analyse its content, and then and automatically assign relevant tags, such as "Positive Review".

There are many approaches to automate the text classification, which are broadly grouped into the following types :

(i) Rule-based – These approaches classify text into organised groups by using a set of handcrafted linguistic rules.

(ii) Machine Learning based - Instead of relying on manually crafted rules, text classification with machine learning learns to make classifications based on past observations. By using pre-labelled examples as training data, a machine learning algorithm can learn the different associations between pieces of text and that a particular output (i.e. tags) is expected for a particular input (i.e. text).

(iii) Hybrid systems combine a base classifier trained with machine learning and a rule-based system, which is used to improve the results further. These hybrid systems can be easily fine-tuned by adding specific rules for those conflicting tags that the base classifier hasn't correctly modelled.

Many of the early works related to text classification were based on "Bag of Words" (BOW) representation, which only accounts for term frequency in the documents and ignores important semantic relationships between key terms. To overcome this problem, some researchers have attempted to enrich text representation by means of manual intervention. However, considering the laborious and time-consuming process involved in manual enrichment of text, few researchers experimented with the use of Wikipedia's knowledge to automatically construct a thesaurus of concepts, which explicitly derives concept relationships based on the profuse structural knowledge of Wikipedia, including synonymy, polysemy, hyponymy, and associative relations [17]. The generated thesaurus serves as a controlled vocabulary that bridges the variety of terminologies present in the corpus of documents. It facilitates the integration of the rich knowledge of Wikipedia into text documents by resolving synonyms and introducing more general and associative concepts, which assist the identification of related topics among text documents. This in turn, facilitates the classification of documents in a better way.

The other prominent use case of Wikipedia has been the classification of small text. With the widespread usage of the internet and consequently many digital touch points, a lot of data about the internet users is generated in the form of web search snippets, forum, chat messages, customer reviews, etc. When it comes to classifying such short texts, there is not enough word co-occurance or shared context to achieve high accuracy. Although some preprocessing technologies such as removing stop words and stemming are proposed to improve the performance, normal machine learning methods usually fail to achieve the desired accuracy due to the data sparseness and background knowledge. In their attempt to improve the accuracy of classification, Xiang Wang et al., mapped the short text to Wikipedia concepts and the concepts, in turn were used to represent documents for text categorisation[18]. After completing this process, traditional classification methods such as SVM, outperformed the traditional BoW approach.

D. Information Retrieval

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)[19]. IR is the process by which a collection of data is represented, stored, and searched for knowledge discovery as a response to a user request (query). This process entails several steps starting with representing data and ending with returning relevant information to the user. The intermediate stage includes filtering, searching, matching and ranking operations. The primary goal of IR is to find the relevant information or a document that satisfies the user's information needs.

Wikipedia's knowledge base has been used in numerous ways to improve the IR process. The IR system developed by Liu, which incorporated the word sense disambiguation algorithm and expanded queries using Wikipedia and WordNet dictionaries, showed an increase in the performance of the retrieval in terms of recall, precision, mean and geometric mean average precisions [20]

ESTER is another efficient search engine that works based on a combination of full text and ontology search [21]. The search engine was built based on Wikipedia and YAGO ontology to process various complex queries in a fraction of a second.

In a similar line of research, Vechtomova proposed new models for retrieving blog posts containing opinions about an entity expressed in the query by building a number of faceted queries (disjunctions of a list of short queries) using Wikipedia.

E. Information Extraction (IE)

Information Extraction is the task of automatically extracting information or facts from unstructured or semi-structured documents[10]. It usually serves as a starting point for other text mining algorithms [9]. For example, extracting entities and their relations from the text can give us useful semantic information.

The IE process takes texts as input and produces fixed format, unambiguous data as output. This data may be used directly for a display to users, stored in a database or spreadsheet for later analysis, or used for indexing purposes in information retrieval applications such as internet search engines like Google [22].

The process of information extraction (IE) turns the unstructured information embedded in texts into structured data. Typical sub-tasks of information extraction include - named entity recognition, coreference resolution and relationship extraction [23].

While few studies have focused on IE from Wikipedia itself, many studies have focused on using Wikipedia's data to improve the IE tasks from other textual sources. It has been observed that several attempts have been made in recent times to use Wikipedia to extract structured information from textual files like HTML. XML etc.

Several researchers have used Wikipedia for Named Entity Recognition (NER). The team from the University of Economics, Prague, has developed NER systems based on Wikipedia's Search API and Apache Lucene search API [24]. Further, there are many popular frameworks for entity linking using Wikipedia, like Dexter, Babelfy etc. Dexter is an open-source entity-linking framework developed by researchers at ISTI-CNR, Italy; Dexter identifies text fragments in a document referring to entities present in Wikipedia. Bebelfy is a multilingual open-source framework with a web interface and an API to perform entity-linking and word sense disambiguation.

Another crucial aspect of text analytics is coreference resolution. Coreference resolution finds the referents of expressions such as pronouns, demonstratives, or definite descriptions within a single document. On a collection of documents, cross-document coreference finds the sets of mentions for each distinct entity mentioned in the collection. Cross-document coreference is not only a useful output of information extraction in itself, but it also supports other information extraction tasks. The coreference resolution also plays a pivotal role in knowledge base construction and is useful for joint inference with other NLP components.

The training and testing of cross-document coreference requires a large labelled dataset and obtaining such a large scale organic labelled dataset is very difficult. One prominent solution to this problem is Wikilinks.

While the extraction of entities like persons and organisations is essential because they form the most basic unit of the information, relations between them play a key role in natural language processing and a better understanding of the text. Relation extraction involves identifying the links between named entities and deciding which ones are meaningful for the concrete application or problem. Given two entities, the aim is at locating the occurrence of a specific relationship type between them. Many researchers in relation extraction tasks have used Wikipedia as a knowledge base. One of the prominent works in this direction is minimally-supervised extraction of domain-specific part-whole relations using Wikipedia as a knowledge base by Ashwin Ittooa and Gosse Bouma. The crux of their approach lies in applying a minimally-supervised algorithm to a large, broad-coverage corpus, which they used as a knowledge base. From this knowledge base, a set of patterns were acquired that reliably express part-whole relations. Further, all triples consisting of the acquired patterns and the instance pairs they connect were extracted from a domain-specific text collection. These triples established the domain-specific part-whole relations.

F. Building Semantic Knowledge Bases

To effectively use the knowledge concealed in Wikipedia, several attempts have been made to extract and transform the unstructured and semi-structured data of Wikipedia into structured and semantically enriched knowledge bases.

A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems, which were the first knowledge-based systems. In recent years, several noteworthy large, cross-domain, and openly available knowledge bases have been created. These include DBpedia [25], Wikidata [26], YAGO[27]. These knowledge bases are not only enabling effective use of Wikipedia's concealed knowledge by humans, but they are also facilitating the better and faster interpretation of knowledge by machines as well. This has resulted in the development of many smart applications, including Question-Answering Systems.

The Knowledge Bases built using Wikipedia have employed different approaches for data curation and storage. The retrieval mechanisms facilitated by these Knowledge Bases are also different. Further, they also differ in their depth and breadth of knowledge.

VI. CONCLUSION AND FUTURE WORK

While the use of Wikipedia for scholarly activities is a subject of debate, Wikipedia is among the top 10 websites in terms of worldwide popularity/usage (Alexa). It is being used extensively by all types of web users, as an easily accessible tertiary source of information about anything and everything, as a quick "ready reference", to get a sense of a concept or idea("Wikipedia," 2019). However, the limited search mechanism provided on Wikipedia and the form in which its data is stored brings in many limitations for using Wikipedia by users and direct interpretation by machines. Despite these lacunae, Wikipedia continues to entice the research community because its data is comprehensive and has a well-formed structure and hierarchical categorisation.

This study found that the collaborative knowledge base of Wikipedia has been used in recent years extensively for different processes and applications related to Text Analytics. The wiki annotations or Wikification has been used to aid the text enrichment and facilitate a better understanding of the text. When used as the background information, it was found that Wikipedia can play a pivotal role in the summarisation of text documents. The content organisation and rich semantic elements of Wikipedia have helped in several studies to improve the Information Retrieval and Information Extraction tasks significantly. The open nature of Wikipedia and its constantly updated, well-formed information has been used to create ontologies. Consequently, it has backed many contemporary knowledge systems such as automated answering, speech recognition, etc.

REFERENCES

[1] H. Chen, R. H. L. Chiang, and V. C. Storey, ‘Business Intelligence and Analytics: From Big Data to Big Impact’, *MIS Q.*, vol. 36, no. 4, pp. 1165–1188, 2012, doi: 10.2307/41703503.

[2] D. Sarkar, *Text Analytics with Python*. Apress, 2016.

[3] R. Feldman and I. Dagan, ‘Knowledge Discovery in Textual Databases (KDT)’, in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Montréal, Québec, Canada, 1995, pp. 112–117. Accessed: Sep. 25, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?id=3001335.3001354

[4] A. Hotho, A. Nurnberger, G. Paaß, and S. Augustin, ‘A Brief Survey of Text Mining’, p. 37.

[5] A. K. Uysal and S. Gunal, ‘The impact of preprocessing on text classification’, *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, Jan. 2014, doi: 10.1016/j.ipm.2013.08.006.

[6] ‘A Survey of Information Retrieval and Filtering Methods’. https://drum.lib.umd.edu/handle/1903/436 (accessed Sep. 26, 2018).

[7] E. D. Liddy, ‘Natural Language Processing’, in *Encyclopedia of Library and Information Science*, 2nd ed., Marcel Decker, Inc., 2001, p. 15.

[8] D. R. Radev, E. Hovy, and K. McKeown, ‘Introduction to the Special Issue on Summarization’, *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, Dec. 2002, doi: 10.1162/089120102762671927.

[9] M. Allahyari *et al.*, ‘A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques’, *ArXiv170702919 Cs*, Jul. 2017, Accessed: Sep. 26, 2018. [Online]. Available: http://arxiv.org/abs/1707.02919

[10] J. Cowie and W. Lehnert, ‘Information Extraction’, *Commun ACM*, vol. 39, no. 1, pp. 80–91, Jan. 1996, doi: 10.1145/234173.234209.

[11] ‘An introduction to Machine Learning’, *GeeksforGeeks*, Aug. 24, 2017. https://www.geeksforgeeks.org/introduction-machine-learning/ (accessed Mar. 24, 2020).

[12] ‘What is Text Classification?’, *MonkeyLearn*. https://monkeylearn.com/what-is-text-classification (accessed Mar. 24, 2020).

[13] R. Mihalcea and A. Csomai, ‘Wikify!: linking documents to encyclopedic knowledge’, in *In CIKM ’07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 233–242.

[14] K. Ramanathan, Y. Sankarasubramaniam, N. Mathur, and A. Gupta, ‘Document summarization using Wikipedia’, in *Proceedings of the first international conference on intelligent human computer interaction*, 2009, pp. 254–260.

[15] ‘Summarizing definition from Wikipedia | Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1’. https://dl.acm.org/doi/abs/10.5555/1687878.1687908 (accessed Apr. 02, 2020).

[16] M. Pourvali and P. D. M. S. Abadeh, ‘A new graph based text segmentation using Wikipedia for automatic text summarization’, *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 3, no. 1, 2012.

[17] P. Wang, J. Hu, H.-J. Zeng, and Z. Chen, ‘Using Wikipedia knowledge to improve text classification’, *Knowl. Inf. Syst.*, vol. 19, no. 3, pp. 265–281, 2009.

[18] X. Wang, R. Chen, Y. Jia, and B. Zhou, ‘Short Text Classification Using Wikipedia Concept Based Document Representation’, in *2013 International Conference on Information Technology and Applications*, Chengdu, China, Nov. 2013, pp. 471–474. doi: 10.1109/ITA.2013.114.

[19] ‘Introduction to Information Retrieval’. https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html (accessed Mar. 31, 2020).

[20] ‘Improve text retrieval effectiveness and robustness - ProQuest’. https://search.proquest.com/docview/304950228 (accessed Mar. 31, 2020).

[21] ‘ESTER | Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval’. https://dl.acm.org/doi/abs/10.1145/1277741.1277856 (accessed Mar. 31, 2020).

[22] J. Davies, R. Studer, and P. Warren, *Semantic Web Technologies*. Wiley, 2012.

[23] L. Liu, Z. Xu, H. Cai, L. Diao, and S. Yan, ‘Acquiring semantic relation pattern from large microblog text’, in *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Aug. 2014, pp. 633–637. doi: 10.1109/FSKD.2014.6980908.

[24] A. Sharma, ‘Exploring Named-Entity Recognition With Wikipedia’, *Analytics India Magazine*, Aug. 24, 2018. https://analyticsindiamag.com/exploring-named-entity-recognition-with-wikipedia/ (accessed Mar. 31, 2020).

[25] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, ‘Dbpedia: A nucleus for a web of open data’, in *The semantic web*, Springer, 2007, pp. 722–735.

[26] D. Vrandečić and M. Krötzsch, ‘Wikidata: a free collaborative knowledgebase’, *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[27] F. M. Suchanek, G. Kasneci, and G. Weikum, ‘Yago: a core of semantic knowledge’, in *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, May 2007, pp. 697–706. doi: 10.1145/1242572.1242667.