International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 11 | Nov -2017 www.irjet.net p-ISSN: 2395-0072

Exploiting Wikipedia and Twitter for Text Mining Applications

D. Jayachitra[1], T.Puvaneswaran[2]

[1]*Assistant Professor, Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli.*
[2]*Research Scholar, Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli.*

Abstract - *Recent research efforts have begun exploring the role of knowledge bases in solving the various problems that arise in the domain of text mining. Of all the knowledge bases, Wikipedia on account of being one of the largest human-curate, online encyclopedias has proven to be one of the most valuable resources in dealing with various problems in the domain of text mining. However, previous Wikipedia-based research efforts have not taken both Wikipedia categories and Wikipedia articles together as a source of information. This research work serves as a first step in eliminating this gap and it has shown the effectiveness of Wikipedia category article structure for various text mining tasks. Wikipedia categories are organized in a taxonomical manner serving as semantic tags for Wikipedia articles and this provides a strong abstraction and expressive mode of knowledge representation. In this research work, it explore the effectiveness of this mode of Wikipedia's expression (i.e., the category article structure) via its application in the domains of text classification, subjectivity analysis (via a notion of \perspective" in news search), and keyword extraction. The effectiveness of exploiting Wikipedia for two classification tasks i.e., 1-classifying the tweets being relevant / irrelevant to an entity or brand, 2-classifying the tweets into different topical dimensions such as tweets related with workplace, innovation, etc. To do so, it defines the notion of relatedness between the text in tweet and the information embedded within the Wikipedia category-article structure. These experimental evaluations undertake comparisons with standard text mining approaches in the literature and the Wikipedia framework based on its category-article structure outperforms the standard text mining techniques.*

Key Words*: *Text mining, feature extraction, supervised learning, Wikipedia and Data mining.

1.INTRODUCTION

1.1 Textual Data over the World Wide Web

Textual data is a very popular means of communication over the World Wide Web in the form of data on online news websites, social networks, emails, governmental websites, etc. Basically, nearly everything which is present on the World Wide Web has a textual presence. In particular, users over social networks generate their own content and prefer to communicate mostly through text.

Textual data has the ability to reach out to a large community, and whenever textual content is read, it can generate a further discussion thereby leading to further generation of textual content. With so much textual data around us especially on the World Wide Web, there is a motivation to understand the meaning of the data through automated methods for all sorts of computer science applications. By understanding the meaning of textual data the machine can answer different questions such as the following:

- What is the main topic and sub-topics of the written text?
- What are the keywords and entities defining the topics of the text piece?
- What is the underlying context of a certain text piece?

1.2 Text Mining

The term \text mining" was first coined in by Feldman and Dagan in 1995. It is the process by which textual data is analysed in order to derive high quality information on the basis of patterns. In the context of text mining, there are two popular classes of techniques namely unsupervised learning and supervised learning. It present a brief overview of each in the following subsections. The last subsection covers evaluation measures used to measure the performance of various tasks.

1.3 Role of Knowledge Bases in Text Mining Applications

Knowledge bases are playing an increasingly important role in solving the various problems that arise in the domain of text mining. Table 1.1 lists a few of the problems along with the knowledge base used to deal with the problem.

1.4 Open Challenges

Despite the application of Wikipedia to several text mining problems, there remain a number of open challenges. It list a few of these challenges

Wikipedia is composed of category hierarchies with the categories linked to Wikipedia articles. To the best of our knowledge, previous research efforts that utilise Wikipedia for knowledge extraction tasks have not taken both Wikipedia categories and Wikipedia articles together as a source of information.

There are several occasions when textual data lacks context and more so in the age of social media. This brings a whole set of new challenges to traditional fundamental research topics in text mining, such as text clustering, text classification, information extraction, and sentiment analysis; unlike standard textual data which has several sentences and hence, a surrounding context whereas social media messages consist of few phrases or sentences. These messages lack sufficient context information for effective similarity measures, the basis of many text processing methods. In such a scenario, external knowledge bases such as Wikipedia can help alleviate the semantic gap in textual data (i.e., lack of context problem).

1.5 Motivation and Problem Statement

Among the fundamental forms of communication, a popular form is written text or textual data. Human beings have found a great comfort in expressing their viewpoint in writing because of the ability to preserve thoughts for a longer period of time than oral communication. However, textual data may contain the following complexities [2]:

- Lack of contextual and background information
- Ambiguity due to more than one possible interpretation of the meaning of text
- Focus and assertions on multiple topics

The above-mentioned problems mainly arise from the informal nature of day-to-day communications of human beings. However, to be able to automatically process textual data, there is a clear need for effective solutions to the above-mentioned issues.

2. Literature Review

In order to cover the related background, it starts by giving a review of different text mining models that are related to our proposed research work. The essential component of any text mining process is conversion of input data from its raw format to a structured, easy-to-manipulate format, a document representation, and it begin by presenting an overview of the \vector space model".

This is followed by an overview of supervised and unsupervised methods of making inferences from textual data. It then briefly present various types of knowledge bases along with a detailed background on Wikipedia which is the knowledge base upon which the contribution of this paper rests. It also motivates our choice of Wikipedia for the text mining applications carried out in this paper . Finally, it present an overview of the microblogging platform Twitter which represents one of the application scenarios to which it apply our semantic relatedness model.

2.1 Unsupervised Learning Methods from Text Data

Unsupervised learning is the name given to the process of finding latent structure in unlabelled data; no supervision implies that there is no human expert who has assigned text documents to classes. The two main unsupervised learning methods commonly used in the context of textual data are clustering and topic modelling. The essential difference lies in whether the membership of a document lies in one cluster (referred to as hard clustering), or in several clusters (referred to as soft clustering).

2.2 Supervised Learning Methods from Text Data

Supervised learning methods are a category of methods that exploit training data (i.e., pairs of input data points with a label for the corresponding output point). These methods learn a classifier or regression function that can be used to compute predictions on new, unseen data. Generally, supervised learning methods for text data fall under the domain of text classification: Figure 2.3 shows an illustration of the text classification process. Some key methods commonly used for text classification are decision trees, rule-based classifiers, linear classifiers, neural network classifiers and Bayesian classifiers.

2.3 Semantic Relatedness

The literature has defined semantic relatedness as a means to allow computers to reason about written text [21] whereby the reasoning deals with finding and quantifying the strength of semantic association between textual units [8]. Within the proposed works in the literature the difference lies in the knowledge base employed, the technique used for measurement of semantic distances and the application domain [9, 10, 12, 16]. Within the context of this paper , it follow the notion of semantic relatedness adopted by Milne and Witten [17] whereby it use it for measuring degree of similarity, and the relationship between different terms.

Two examples from Milne and Witten are with respect to relationship between \social networks" and \privacy", and \cars" and \global warming". To clarify further, `lion' and `cheetah' are not same but are similar due to belonging to the same biological family i.e., Fieldale; likewise word pairs carpenter: wood and mason: stone are relationally similar because both carpenter and mason are professions while wood and stone represent materials used to carry out the job. Note that it utilise Milne and Witten's definition of semantic relatedness; however, it differ with them in terms of strategy employed since they utilize Wikipedia hyperlinks which in our context fails to show good performance1. For the contributions in this paper , the semantic relatedness framework introduced in Paper 4 is a core component, and here it present some semantic relatedness frameworks proposed in the literature.

Table 3.1: Example showing application of NER over a sentence

Type	Sentence
Input(Annotated)	Joe and Alan worked for Luther Corp. in 1982.
Output(Annotated)	[Joe]PERSON and [Alan]PERSON worked for [Luther Corp.]*ORGANIZATION* in [1982]*TIME*.

Relatedness measures there is inconsistency in results, and one underlying reason for this is the different application scenarios for which they have been devised. It differ from proposed techniques in that it utilise Wikipedia categories in conjunction with Wikipedia articles whereas earlier works utilise either Wikipedia hyperlinks or category hierarchies without taking into account their combination.

3.2 Named Entity Recognition

Named Entity Recognition (NER) is a key task in information extraction and forms related work for our contributions made in Paper 5. NER fundamentally involves annotating a snippet of text with a label from a set of fixed category types such as name of person, location, quantities, percentages, products, time etc. Formally, NER is performed in two steps: first, different block of texts are extracted from a document and then, each block of text is classified into a different range of category types. Table 3.1 shows the application of NER over a sentence, where upper-case words show the annotated category type for the block of text by NER.

3.3 Knowledge Extraction

Knowledge extraction aims to preserve the meaning of textual units of information by providing a concise representation of documents. Specifically, it consists of approaches for document summarization and keyword extraction. However, due to it being a closely investigated area related to keyword extraction, it include it as a related work in the paper for a better presentation of the related research. First, it present an overview of document summarization; it is the task that generates a summary of a document in a few words while retaining the important points of the document. Then, it present an overview of keyword extraction; it is the task that extracts most important keywords which represent the gist of a document while omitting the sentence based structure of the document.

4. Proposed Approach

This section presents a brief overview of our approach for the filtering task and the reputation dimensions classification task. Fundamentally, the approach is aimed at enhanced context representation for tweets in order to filter them with respect to entities and/or reputation dimensions; this is done in an effort to address the second research question raised in Section 1.3 (Paper 1). The strength of our approach consists of the exploitation of the encyclopaedic knowledge in Wikipedia which is an up-to-date and dynamic resource with extensive knowledge on various subjects as explained in Paper 2.

4.1 Filtering Task

The task of filtering tweets is performed through supervised learning by training the classifier using the following feature types:

- Relatedness scores for several (entity-related) Wikipedia category taxonomies
- Topical scores corresponding to each tweet obtained via topic modeling
- Twitter-specific features obtained using the Twitter API

The fundamental constituent of the technique is the Wikipedia-based features which make use of the Wikipedia category-article structure that describes the entity to obtain a suitable set of related terms corresponding to an entity.

It also experimented with one such approach with details in Appendix B; our system based on Wikipedia hyperlinks does not exhibit optimal performance whereas the one based on Wikipedia category article structure that it explain in this shows a performance comparable to the one exposed by the best systems participating in the filtering task.

4.2 Reputation Dimensions' Classification Task

The task of reputation dimensions classification is performed through supervised learning by training the classifier using the following feature types:

- Relatedness scores for several (reputation classes related) Wikipedia category taxonomies
- Statistical features which it further categorize into tweet-specific features, language specific features, and word-occurrence features described in the following.

The fundamental constituent of the technique is the Wikipedia-based features which make use of the Wikipedia category-article structure that describes a reputation dimension to obtain a suitable set of related terms corresponding to that dimension.

4.3 Methodology

In this section it present the proposed methodology that it have defined for the tasks of filtering and reputation dimensions' classification.

Psuedo-Code for selecting category taxonomies

```
def main():
 get_selectedTaxonomies(entity_Wiki_article) # main
    functionality

def generate_taxonomy(rootcat, d):
 return (sub_cat(cat, 2), rootcat) # returns sub categories
    along with root

def get_lst_taxonomies(relscores):
 taxonomies = []
 for sc, taxonomy in relscores:
  taxonomies.append(taxonomy)
 return taxonomies

def get_selectedTaxonomies(wiki_article):
 selected_taxonomies = []
 for pcat in categories(wiki_article):
  selected_taxonomies.append(generate_taxonomy(pcat))

 D = Merge all tweets belonging to single domain/entity #
    one big document of tweets
 phrases = extractVariablePhrases(D) // Psuedo-code
    presented in Listing 4.1
 cat_lst = set()
 for p in phrases:
  cat_lst.union_update(getCategories(p)) // since each
     phrase is a wiki article
 relscores = []
 for cat in cat_lst:
   cat_t = generate_taxonomy(cat, 2)
   score = relatedness(D, cat_t)
   relscores.append([score, cat_t])

 ordered_relscores = relscores order by score (descending
    order)
 additional_taxonomies = get_lst_taxonomies(
    ordered_relscores[:100]) # select top-k taxonomies
 selected_taxonomies.extend(additional_taxonomies)
 return selected_taxonomies
```

4.4 Feature Set Based on Topic Modelling

A well-known topic modelling technique known as Latent Dirichlet Allocation (LDA for short) [24] is used for this set of features. LDA is an unsupervised machine learning technique aimed at identification of latent topics in large document collections. It is built on the \bag of words" approach with each document being treated as a vector of word counts and finally as an outcome of LDA, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. It trained LDA with 300 topics on each domain (music, automobile, etc) containing several entities, and the score (i.e., probability distribution) in each topic is then utilised as a feature, and hence all topics make a feature set. The rationale for this is that the Wikipedia article titles cannot match all the terms and therefore, with the help of LDA it can include the inuence of the remaining terms.

4.5 Twitter-Specific Feature Set

In this section it present the set of features that are specific to the nature of Twitter. It categorize these features into three categories: tweet content features, user information features, and mention features.

Tweet content features: These are features derived from the content of tweets.

User information features: These are features derived from the profile information of the Twitter user who is the producer of the tweet.

Mention features: These are features derived from the profile information of the users that are mentioned in a tweet.

Table 5.2 shows the detailed description of these features. Note that the profile information features for the users who produce a tweet or are mentioned in a tweet are utilised only in cases when the profile information is available (i.e., in cases where the user profile is public and has not been deleted or blocked from Twitter). The use of Twitter-specific features helps enriching the machine learning model which in turn improves the classification accuracy.

This is on account of specific attributes of Twitter whereby organizations and individuals use it differently. Moreover, each entity differs from the other in terms of its presence on Twitter; as an example certain Spanish banks from within the dataset have an overly active Twitter presence due to their sponsorship of football clubs. Note that Table 5.2 to the best of our knowledge shows an exhaustive set of Twitter-specific features and the selection of these features was motivated by standard works on tweet classification from within the literature [20, 52, 158, 165, 176, 197, 223].

4.6 Twitter Dataset

It use the dataset provided by CLEF 2013 RepLab task organizers which is a multilingual collection of tweets (i.e., 20.3% Spanish tweets and 79.7% English tweets). The corpus contains tweets referring to a set of 61 entities from four domains; automotive, banking, university, and music. The filtering task utilised tweets from all four domains whereas the reputation dimensions' classification task utilised tweets from automotive and banking domain.

Table 4.1: RepLab 2013 Dataset Details

	All	Automotives	Banking	University	Music
Entities	61	20	11	10	20
Training No. Tweets	45,679	15,123	7,774	6,960	15,822
Test No. Tweets	96,848	31,785	16,621	14,944	33,498
No. Tweets EN	113,544	38,614	16,305	20,342	38,283
No. Tweets ES	28,983	8,294	8,090	1,562	11,037

The tweets were gathered by organizers of the task by issuing the entity's name as the query. For each entity roughly 2300 tweets were collected with the first 750 constituting the training set, and the rest serving as the test set. Table 5.3 shows the statistics of the dataset.

4.7 Wikipedia

The data for Wikipedia category-article structure is obtained through a custom Wikipedia API that has pre-indexed Wikipedia data and hence, it is computationally fast15. The API has been developed using the DBPedia [22] dumps and it is a programmer-friendly API enabling developers and researchers to mine the huge amount of knowledge encoded within the Wikipedia structure.

Here, it show that the entity filtering task can be effectively addressed by approaches relying on Wikipedia; in fact the new enhanced approach to the filtering task shows a performance comparable to the one exposed by the best systems at RepLab 2013 as it will show in the evaluations reported in this chapter.

5. Experimental Results for Reputation Dimensions' Classification Task

Table 5.2 presents a snapshot of the official results for the filtering task of RepLab 2014, where CIRGIRDISCO is the name of our team. As can be seen from Table 5.2, out of a total of 8 participating teams in RepLab2014 reputation dimension classification task 4 teams outperform our best run. Our system shows good results for the evaluation measure of accuracy; however, the evaluation measures of precision and recall show an average performance and one reason for this is due to our training

Table 5.1: Comparison of Experimental Results for Systems in CLEF RepLab 2013 Task

Setting	Sensitivity	Reliability	F-Measure
Our Approach	0.4450	0.8351	0.4870
Baseline	0.4902	0.3200	0.3255
POPSTAR [184]	0.7288	0.4507	0.4885
SZTE [86]	0.5990	0.4444	0.4385
Previous Approach [171]	0.4164	0.6687	0.4485

Table 5.2: Results of Reputation Dimensions' Classification Task of RepLab 2014

Team	Accuracy	F-measure
uogTr_RD_4	0.7318	0.4735
DAE_RD_1	0.7231	0.3906
Lys_RD_1	0.7167	0.4774
SIBTEX_RD_1	0.7073	0.4057
CIRGIRDISCO_RD_3	0.7071	0.3012
CIRGIRDISCO_RD_2	0.6924	0.2386
Baseline	0.6222	0.4072
CIRGIRDISCO_RD_1	0.6073	0.3195

Table 5.3: Proportion of Relevant and Irrelevant Tweets for Some Entities inTraining Data

Entity (Domain)	Total Training Tweets	Irrelevant Tweets	Relevant Tweets
Adele (Music)	694	20	674
Jennifer Lopez (Music)	862	4	858
Led Zeppelin (Music)	908	0	908
Maroon 5 (Music)	738	0	738
Bankia (Banking)	760	19	741
Barclays (Banking)	747	1	746
HSBC (Banking)	797	6	791

Table 5.4: Proportion of Relevant and Irrelevant Tweets for Some Entities inTraining Data

Sample	Innovation	Citizenship	Leadership	Workplace	Governance	Undefined	Performance	Products and Services
Training Data	313	2209	297	468	1303	2228	947	7898
Test Data	306	5027	744	1124	3395	4349	1598	15903

and testing methods being applied over eight classes because it included the class \Undefined" in our training and testing supervised learning method whereas the RepLab 2014 organizers excluded this class. However, it was not clear in the task guidelines.

In summary, classifying tweets into relevant or irrelevant for an entity or along various reputation dimensions is a challenging task with most of the challenges stemming from the nature of how text is written by Twitter users. In this section, it perform an analysis of our proposed methodology in an attempt to perform a detailed study of the effectiveness of the proposed features

Table 5.5: Standard Deviation of F-MeasureR, F-MeasureI, and F-MeasureRSfor Various Domains in Dataset

Domain	F-Measure$_R$	F-Measure$_I$	F-Measure$_{RS}$
Automotives	0.1115	0.3180	0.2746
Banking	0.1230	0.4105	0.3767
University	0.1689	0.2971	0.2169
Music	0.0391	0.4058	0.3974

6. Conclusion and Future work

There is a need for advances in algorithmic design which can learn interesting patterns from textual data. Wikipedia categories are organized in a taxonomical manner serving as semantic tags for Wikipedia articles and this provides a strong abstraction and expressive mode of knowledge representation. It used this mode (i.e., Wikipedia's category-article structure) in the domains of text classification, analysis (via a notion of \perspective" in news search), and keyword extraction. For text classification and subjectivity analysis, it have proposed a semantic relatedness framework which first extracted phrases matching Wikipedia article titles/redirects, and then utilised these phrases in matched Wikipedia categories corresponding to the entity of interest in order to determine the relatedness between phrases and the entity of interest.

As with any human-curated effort Wikipedia despite its wide-scale coverage of knowledge has some limitations which affect the outcomes of this paper. The phrase chunking strategy introduced in this paper may have tendency to miss out significant phrases on account of Wikipedia missing out some information on long-tail entities.

References

[1] SisayFissahaAdafre and Maarten De Rijke.Finding similar sentences across multiple languages in wikipedia.In Proceedings of the 11th Conference of the European paper of the Association for Computational Linguistics, pages 62{69, 2006.

[2] Charu C Aggarwal and ChengXiangZhai. Mining text data.Springer Science & Business Media, 2012.

[3] Eugene Agichtein, Eric Brill, and Susan Dumais.Improving web search ranking by incorporating user behavior information.In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '06.ACM.

[4] EnekoAgirre and AitorSoroa.Personalizing pagerank for word sense disambiguation. In Proceedings of the 12th Conference of the European Paper of the Association for Computational Linguistics, EACL '09, pages 33{41, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[5] RakeshAgrawal, SreenivasGollapudi, Alan Halverson, and Samuel Ieong.Diversifying search results.In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pages 5{14.ACM, 2009.

[6] Rodrigo Aldecoa and Ignacio Marfifin. Exploring the limits of community detection strategies in complex networks. Scientific reports, 3, 2013.

[7] Enrique Alfonseca and Suresh Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In Proceedings of the 1st international conference on general WordNet, Mysore, India, pages 34{43, 2002.

[8] Enrique Amigo, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martin, Edgar Meij, Maarten de Rijke, and DamianoSpina. Overview of replab 2013: Evaluating online reputation monitoring systems. In CLEF 2013 Labs and Workshop Notebook Papers, Springer LNCS, 2013.

[9] Enrique Amigfio, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and DamianoSpina. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In Information Access Evaluation. Multilinguality, Multimodality, and Interaction, pages 307{322.Springer, 2014.

[10] Enrique Amigfio, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Maarten de Rijke. Overview of replab 2012: Evaluating online reputation management systems. In CLEF 2012 Labs and Workshop Notebook Papers, 2012.

[11] Enrique Amigfio, Julio Gonzalo, and FelisaVerdejo.A General Evaluation Measure for Document Organization Tasks.In Proceedings SIGIR 2013, pages 643{652, July.

[12] JisunAn, Meeyoung Cha, P Krishna Gummadi, and Jon Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In ICWSM, 2011.

[13] S□oren Auer, Christian Bizer, GeorgiKobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. Springer, 2007.

[14] Ricardo Baeza-Yates and Carlos Castillo. Crawling the infinite web: five levels are enough. In In Proceedings of the third Workshop on Web Graphs (WAW), pages 156{167.Springer, 2004.

[15] Breck Baldwin and Thomas S Morton.Dynamic coreference-based summarization.In EMNLP, pages 1{6, 1998.

[16] L Baleydier. The kartoo visual metasearch engine. 2008.

[17] Ken Barker and Nadia Cornacchia. Using noun phrase heads to extract document keyphrases. In Advances in Artificial Intelligence, pages 40{52.Springer, 2000.

[18] Regina Barzilay and Kathleen R McKeown. Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3):297{328, 2005.

[19] S. Bergamaschi, F. Guerra, and B. Leiba. Guest editors' introduction: Information overload. Internet Computing, IEEE, 14(6):10{13, Nov 2010.

[20] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. Broadly improving user classification via communication-based name and location clustering on twitter. In HLT-NAACL, pages 1010{1019, 2013.

[21] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what's in a name. Machine learning, 34(1-3):211{231, 1999.

[22] Christian Bizer, Jens Lehmann, GeorgiKobilarov, S□oren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann.Dbpedia - a crystallization point for the web of data. Web Semant., 7(3):154{165, September 2009.

[23] Christian Bizer, Jens Lehmann, GeorgiKobilarov, S□oren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann.Dbpedia-a crystalliza-tion point for the web of data. Web Semantics: science, services and agents onthe world wide web, 2009.

[24] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 2003.