International Journal of Engineering & Technology, 7 (4.19) (2018) 609-614

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Enriching Tweets for Topic Modeling via Linking to the Wikipedia

Ghaidaa A. Al-Sultany*[1], Hiba J. Aleqabie[2]

[1]*Department of Information Network, College of Information Technology, University of Babylon, Babel, Iraq*
[2]*Department of Software, College of Information Technology, University of Babylon, Babel, Iraq*
**Corresponsing Author E-mail: ghaidaa.almulla@it-net.uobabylon.edu.iq*

Abstract

Twitter is currently the essential broadcasting services in the world of multimedia and the scope of spread information, news, and events. Due to the nature of tweets, from the short post, limited contextual information, sporadic, noisy and vague, the topics of learning remain a significant challenge. In this paper, the proposed approach overcomes those challenges through complete the contextual information of each tweet by attaching descriptions from the Wikipedia. Tweets will be enriched and integrated with its contextual information of the Wikipedia, in order to convert short posts into long texts and get rid of the deficiencies that produce during the training. This procedure implemented by following the steps; firstly, Twitter Name Entity Recognition (TNER) was proposed to classify the tweets by the nature of it and choosing specified entities for the next step. Secondly, establish connecting through the Wikipedia API, linking the entities of each tweet and Wikipedia appending it to its tweets creating a new dataset, feeding the preprocessing step. Finally, topic modeling, Latent Dirichlet Allocation (LDA), was applied.

Moreover, a comparison based on the effect on modeling representation, and the nature of the topics for both datasets was performed. As well as the evaluation criteria were performed i.e. perplexity and coherency of both models.

The twitter Dataset was collected via API, from several Twitter accounts for Fox News, Reuters, and CNBC. It indicates that the system affects the representation of topics for the topic modeling. The representation was better for enriched tweets, and the tokens of each topic more descriptive and meaningful, this was indicated by the high coherency of the second model that improve and affect the representation of topics.

Keywords: *Twitter, Topic modeling, Wikipedia, TNER, NER, and LDA.*

1. Introduction

Since its initiate in 2006, Twitter has turned into a vast phenomenon while it was a comprehensive service's tool. It is spread through the globe successfully; Twitter currently accessible in 33 diverse languages, supporting non-Latin languages' alphabet. [1]. Twitter characterized by short messages; inclusion of URIs; username mentions; topic markers; and threaded conversations. It often presents local content containing abbreviations and errors[2].

Entity linking, the process of identifying the entity indicated in the tweet refers to, provides the ability to link tweets to existing Wikipedia's page. Moreover, thus supports multiple natural language perception. One of the challenges was capturing semantic and background information of entities that have a page at Wikipedia' site. Each entity may have multiple entries to Wikipedia. This challenge was solved by getting the Wikipedia API.[3].

The massive amount of electronic document was analysis by a robust technique that is Topic Modeling. It is utilizing for finding hidden themes that collection. Topic modeling can bind words with similar context and the recognized through various meanings of words. Extracting relations and the expressive information is a great challenge, especially for a large amount of data, and consider as an effective method for discovering hidden structures in those data. Topic modeling is a robust method that performs further than clustering or classification approaches. It is sampling the objects as latent groups (topics) reflecting the content and concepts of the data. Topic modeling applied to diverse texts mining applications, i.e. summarization, document classification[4].

The primary goal was to supplement and enrich the tweets with context-based information. This information was provided by the Wikipedia. The operation was performed by converting the tweet's posts into entities. The entities were linked into its corresponding Wikipedia API. Then Extract the descriptive abstract for each entity. Attach these abstracts to its tweet and construct a new augmented dataset. A standard topic modeling, LDA was performed, and finally, a comparison was made for both datasets using perplexity and coherency models.

The paper was organized as section II is the related works Section III for topic modeling, LDA. Section IV; shows a brief description of Wikipedia. Section V for the proposed system. Section VI for experimental results and discussion. Finally, conclusion and future work in the section in section VII.

2. Related Works

Although conventional topic models have achieved great success for regular-sized documents, they do not work well on short text collections. Since a short text (tweets) only contains a few meaningful keywords, the word co-occurrence information is complicated to be captured[5], and The sparsity of content in short texts brings a new challenge to topic modeling[6]. Recently numerous attempts were devoted to handling this challenge. Some methods had been proposed to expand the representation of short text using latent semantics or unrelated words. Favorite strategies are aggregating short texts to the pseudo-documents and uncover the cross-document word co-occurrence[5]. Information was extended through training datasets locally or from exterior corpus such as Wikipedia. Another approach suggests using the graph-based measurement, for the similarity of text and involving Wikipedia as background knowledge trying to get the semantic likeness through documents[6].

[7] Shows that link analysis of the topic-keywords graph performed the enrichment procedure of short text classification. Re-rank, the keywords' distribution, extracted by a "bitterm topic model" to make the topics further noticeable via constructing the topic-keyword graph and conducting link analysis. [8] Propose WikiLDA as an improvement to LDA employing Wikipedia. The proposed method starts by appending each document in the corpus, to its related Wikipedia concepts. Then use the Generalized Pólya Urn (GPU) to merging word-word, word-concept, and concept-concept semantic relatedness into the generative process of LDA. [9] Propose an entity-topic modeling approach for integrating the DBpedia background knowledge about entities such as the occupation of persons, the location of organizations, and a band of a musician….etc. improve document clustering and yield a semantic topic modeling. [10] Presenting the KEA method, at the "#Microposts 2016 NEEL" Task. Analyze the English microspots and connect the documents entities into it is correspondence entities in DBpedia. NLP tools performed this method.

When natural language processing applies to short text, i.e., Twitter, would not work well due to the nature of tweets' texts. Especially, when dealing with part-of-speech and name entity recognition tweets would harden the operations.[2] Presents a twitter tagger and evaluate methods for improving English part-of-speech tagging implementation. Also, suggest an error analysis method for the current tagger. Motivating a set of tagger augmentations was verified to raise the performance. Moreover, presenting a new approach where available taggers use different tag sets to handle the case of unknown words and slags. [11] Suggest an approach for the CAp 2017 challenge. Tweets' dataset of French tweets was provided, which the first such dataset in French. Suggesting name entity recognition (NER) for tweets, 13 types of entities were presented. [11] Advise a study was made for finding a confidence level to the problem of detecting entities in tweets. The problem was outlined as a binary classification, and for finding the probability of a nominee named entity is a real-named entity, a recurrent neural network was used.

3. Topic Modeling

This strategy is utilized to consequently find out the topics from an accumulation of documents, with the instinct that each document shows different topics. LDA models the document's words were producing a from a combination of topics where every topic is a latent distribution of word probabilities [12].

LDA is a generative model that trains groups of documents into a set of latent topics. The input to LDA was a "bag-of-words" explanation for the single documents, and the output was a group of latent topics. And for each document, several topics were allocated. Officially, a topic was a multinomial appropriation of words, and a document was related with a multinomial appropriation of topics [13]. LDA assumes the following generative process for generating documents.

1. Select word probabilities (ϕ_t) for each topic t: $\phi_t \sim$ Dirichlet(β)
2. Select topic proportions (θ_d) for document d: $\theta_d \sim$ Dirichlet(α)
3. Select the topic for each word position ($z_{d,n}$): $z_{d,n} \sim$ Multinomial(θ_d)

where α and β are Dirichlet priors for document-topics and topic-words distributions respectively[8].

4. Wikipedia

Wikipedia is a multilingual, web-based, free content encyclopedia that Attracts many people. English Wikipedia is the most massive and most well-formed. English Wikipedia by itself possesses greater than 1 billion words, which exceed the amount size of 'the Britannica Encyclopedia,' 25 times [14]. Wikipedia includes over four million articles on many topics. Moreover, According to the Wikipedia strategy, the style of the article imposes that the article's title should be a Name or a depiction of the topic [15]. It includes a large kind of concepts in various areas, i.e., "Arts, Geography, History, Science, Sports, and Games." Since it turns into a database saving all human knowledge, a promising approach that connects the Semantic Web and the Social Web (a. k. a. Web 2.0) called Wikipedia mining was founded. As a corpus for knowledge extraction, Wikipedia's attributes are not restricted to the scale, but also have the dense link structure, sense disambiguation based on URL, brief link texts and well-structured sentences.[16].

5. The Material and Method

The proposed approach is consisting of 6 stages. Constructing twitter dataset, Twitter Name entity recognition (TNER), Wikipedia linking, Enriching tweets, preprocessing, and topic modeling. See Figure 1.

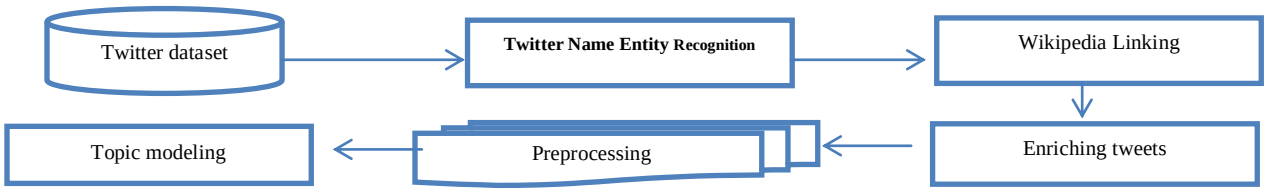

Fig. 1: Proposed System

5.1. Twitter dataset

The dataset was collected via Twitter API. The Twitter API platform offers choices for streaming real-time Tweets. [12][20]. The Streaming API is probably the most broadly used data source for Twitter studies. Regularly, broad-scale quantitative examinations of Twitter information depend on raw data gathered through this source. This stream of data is given only as a live poll, implying that the minute a tweet is posted on Twitter, it became accessible [1]. The stream in JSON file formatted, an extraction of the text from the tweets JSON file was required, See Fig 2.

5.2. Twitter Name Entity Recognition (TNER)

The name entity recognition is the operations performed on texts that look for and detects entities in the text. Classify them into specific categories such as "names of persons, organizations, locations, expressions of times, quantities, monetary values, and percentages."
Due to the structure of tweets; inclusion of URIs; The user's mentions; hashtags; retweets and conversations, the NER became a challenge [15]. This challenge was solved by developing an algorithm that, merely classifies the tweets into its main symbols and consider them as entities: Hashtag, Mention, Url, and RT entities, incorporation with standard NER to extract the general entities from the rest text. See algorithm 1.
To clarify the idea of this procedure a tweet was taken as an example. The tweet will be tokenized into a list of tokens; classification was applied to produce a pair of (word, label). See Table 1.
"@FoxNews NATO allies have financed Turkey. trump is doing Putin's bidding to strike at NATO and destabilize the was… https://t.co/9rleDadLu6".

5.3. Wikipedia linking

```
Algorithm TNER
Input  Tweets
Output Entities

Begin
   For all tweet ∈ Tweets do
     Tokenize tweet
      Case t in tweet:
         - t  start with  'RT'    :  create  pair   (t, 'RT')
         - t  start with  '#'      :  create pair   (t, '#hashtag')
         - t  start with  '@'     :  create pair  (t, '@mention)
         - t  start with  'https' :  create pair   (t, URL')
        - Else :
              go to NER to extract the primary entities.
      End case
  End for
End
```

Algorithm 1: twitter name entity recognition (TNER)

```
Algorithm Wikipedia API establishing
Input  TNER
Output Set of abstracts

Begin
Entities= List of Person, Organization, and Geopolitical
entities' pairs.
For all Entity ∈ Entities do
  Extract the words' pair.
  Apply normalization on that word, treated as a title.
  Establish a connection with the Wikipedia websites.
  Search for an entry in the Wikipedia website for the
word.
  If entry found then
  Extract the description abstract of this entities.
End for
End
```

Algorithm 2: Wikipedia API establishing

After annotating words, each word classified as a person, geographic location and organization were assigned as candidate entities. These entities are then normalized to qualify the way of Wikipedia name strategy; each word starts in capital letter. Wikipedia contains million of the page for probably all the characters, celebrities, icon, cities, countries and Continents. Examine if the candidate entity has a page or link on Wikipedia. If it has the next step would be, accessing that Page, fetches the description of the Wikipedia page. Finally, Store the descriptions in a file besides the tweets. The work was performed online due to the massive amount of storage it needs and the inability to download all Wikipedia also, try to get them all updated occurs on the pages. See Algorithm 2.

Table 1: TNER

Text	Label	Description
@FoxNews	@	Mention to another account
NATO	ORG	Companies, agencies, institutions.
Turkey	GPE	Geopolitical entity, i.e., countries, cities, states.
Putin	Person	People, including fictional.
NATO	ORG	Companies, agencies, institutions.
https://t.co/9rleDadLu6	Url	Https:// link to another tweet

5.4. nriching tweets

After completing the linking process with Wikipedia, each tweet will be attached to a set of descriptors extracted from the previous step. Each tweet has its descriptions and according to the number of words and the quality of the words.

5.5. Preprocessing[17]

The Twitter data required a thorough clean to ensure that identifying valid and representative patterns of topics. The tweets are concise. Probably contains some slags words or abbreviations. Also, a frequent word in the collected dataset is "RT" which is a common word that indicates a retweet. This word is removed where it is already in our stop words.
Tweets may include many URL usually used to give the source for the detailed description of the content mentioned in the tweets. Hence all such web URLs is removed from the tweets by identifying such patterns. User mentions are used commonly in tweets to refer to or mention another person. The mention is usually Start with '@' symbol; this mention removed also. Hashtags are another entity

associated with a tweet to name or tag a topic and usually start with a # symbol. Hashtag need not be a meaningful word. These entities are removed since found no significance in our scoring approach. Exclude all Words with three letter length as they would not be helpful to a topic model. Long words lower the opportunity of appearance for those less frequent words. Excluding the Stopwords, as they are mostly uninformative in a topic model. Also, Words containing non-English characters and "emoji" and slags words were removed. Finally, the part of speech was applied that indicate for each token a label Verb, Noun, Adj, Advs.,…etc., for the proposed system, selecting only the nouns tokens.

5.6. Topic Modelling

After cleaning the texts, the input to the topic modeling is bag-of-words depiction of the each tweet's tokens; two models were used LDA and NMF for topic modeling. Testing the enriching process and how it will affect the topic representation of both models. LDA can only use raw term counts for LDA because it is a probabilistic graphical model.

Table 2: Experiment's amounts

Tweets	3468
words (tokens)	45,398.
Entities	13974
candidate entities	3702
Real entities	2334

An example will be illustrated to explain the working procedure.

The tweet:
"@FoxNews NATO allies have financed Turkey. trump is doing Putin's bidding to strike at NATO and destabilize the was… https://t.co/9rleDadLu6 "

Step 1: TNER
- The tweet's tokens will be classified into the following entities:
  ('@FoxNews,'@')
  ('NATO', 'ORG')
  ('Turkey', 'GPE')
  ('Putin', 'Person')
  ('NATO', 'ORG')
  ('https://t.co/9rleDadLu6','Url')
- Selecting the entities with the labels ORG, GPE, Person.
  ('NATO', 'ORG')
  ('Turkey', 'GPE')
  ('Putin', 'Person')

Step 2: Wikipedia Linking
- Establish a connection with Wikipedia
- Send the entity, i.e. ('NATO)
- If the NATO link to Wikipedia web page
  -Request the descriptive of Nato.
  Else
  -Get the next entity.

The result of this step
NATO ("The North Atlantic Treaty Organization…")
Turkey ("Turkey, officially the Republic of turkey…")
Putin ("Vladimir Vladimirovich Putin is a Russian politician…")

Step 3: Tweet Enrichment
The tweet will be appended by the description for each entity result in step 1, assign them as a single document.

"@FoxNews NATO allies have financed Turkey. trump is doing Putin's bidding to strike at NATO and destabilize the was… https://t.co/9rleDadLu6 " NATO ("The North Atlantic Treaty Organization, …...") Turkey ("Turkey, officially the Republic of turkey,…") Putin ("Vladimir Vladimirovich Putin is a Russian politician …").

Step 4. Applying Preprocessing then Step 5 will be performed to result in the topics shown in Fig 3.

6. Experimental Result and Discussion

The tweets were collected from the Fox News, Reuters, and CNBC twitter accounts'. Once, data collection complete, texts were extracted from tweets that are in JSON formula. The number of tweets was 3468, of words (tokens) 45,389. The number of entities was 13974. The candidate entities were 3702; real entities that have corresponding page 2334, the unique entities were 1008. See Table 2.
Creating the second dataset as a collection of documents resulting from the enrichment process, where each tweet and Its attachment was mapped into a document.
The enrichment process begins with, for each tweet's text; an annotation was performed for each tweet using TNER. Obtaining a

list of pairs (token, annotation) arranged for each tweet. Specified entities were selected the Person, Organization, and Gpe since most of them had pages in Wikipedia. Depending on the nature of the title for the pages in the Wikipedia, the page's title should begin with a capital letter for each word. For this reason, Normalization was performed, converting the token's words into the same page title format. The search process was more comfortable and faster than other entities.

Now, for every word, a page will be searched for in the Wikipedia site related to that word. By seeking for an API and using a particular type of query the Extraction of a descriptive abstract of the page was done. Often this page is stored in JSON Format. Accessing, the JSON field that is important and leaves the rest fields of the page. This field is the summary that displayed on the Wikipedia page.
All the abstractions were attached to its tweet to rebate the ambiguity, lack of meaning, assign them to a document to provide the best representation in topic modeling later.
In the next stage, the preprocessing will be applied. As mentioned in the preprocessing paragraph, the texts of tweets will be cleaned by removing the various symbols mentioned above, all words were converted to lower case, removing the stop words and eliminating each word with some characters is less than 3 and more than 15, and removing punctuations. Part of speech tagging was also applied selecting only the nouns (Treebank NN). This step was essential to building most topic models. Finally, examine these words whether they are a meaningful word or not, excluding all meaningless words, this stage ensures that text mining would identify a valid and representative pattern for topic modeling.
Topic models were regularly evaluated using internal metrics, i.e., held-out probability, perplexity and it is important to look at the topics themselves, and see if they tell a coherent and sensible story. See Table 3. Not just consider the main factors of evolution: Number of topics, the top 10 words, but also consider the distributions of words within and across topics. The proposed model use word clouds to visualize the top N words, and represent words that are more or less informative in a natural style.

The experiment was conducted on several times for both datasets. The number of topics was assigned for the range of values 10, 20, 30, 40, and 50 to assess the achievement of the experiment for a various number of topics. Each value of the number of topics for the LDA algorithm was run with iteration =50, passes=50 and considering the evaluation metrics.

Table 3: Evaluation

Number of topics K	Dataset 1		Dataset 2	
	Coherence	Perplexity	Coherence	Perplexity
10	0.6017	-6.9063	0.6903	-5.0918
20	0.5819	-6.9275	0.6578	-4.8946
30	0.6051	-6.9523	0.6438	-4.8476
40	0.6075	-6.9989	0.6719	-4.7903
50	0.6158	-6.9653	0.6509	-4.7569

Fig. 2: Sample Of Topics For The First Dataset

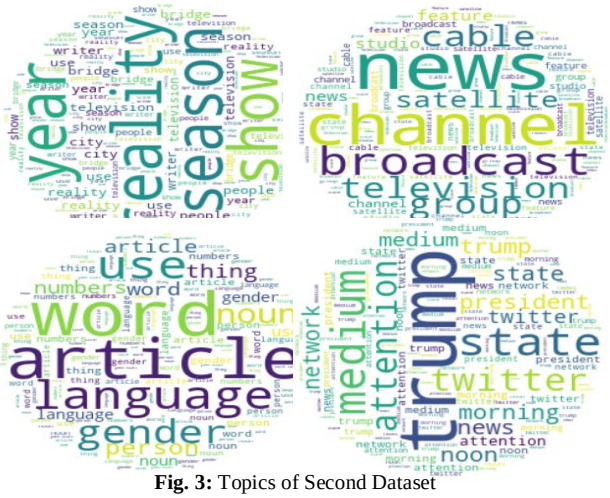

Fig. 3: Topics of Second Dataset

In each execution, and for both datasets, top 10 words, with the highest probability, for each topic were retrieved, where word clouds were used to visualize the top N words and represent words that are more or less informative in a natural style, as illustrated in Figure 2 and Figure3.

It is essential to look at the topics themselves and see if they tell a coherent and sensible story. Not to only considering the probability of the top 10 words, moreover, considering the distributions of words in and through topics. In Figure 2, the topics representing the LDA for the first dataset. Each topic consists of a group of words that are shown in a visual form. Where the size of the word indicates the high value of distribution probability in that topic. While in figure 3, When reading the topics, holding the words starting from the biggest, to smallest, a feeling like a story was told due to the coherent and close meanings words that show, what this topic is about?

The second evaluation as illustrated in Table 3 confirm the perplexity of likelihood quality and model coherency for different values of the number of topics and for both dataset. A good model will generate coherent topics, i.e., topics with the highest topic coherence scores. From Table 3 an understanding was obtained that the importance of coherency between the topics raises topical the representation. For the second dataset, the coherence increase in ranges that +

the first dataset. Furthermore, it will be noticeable to indicate that, the varying values of coherence by the different number of topics. The coherence score looks at which is better sense to pick the model that offered the peak score before pulling down. Alternatively, the perplexity score seems to possess decreasing; it may make better sense to pick the model that donated the lowest score before pulling up.

7. Conclusion and Future Work

The primary challenge in dealing with Twitter was the database. The available dataset was collected in streaming as raw data. Thus our dataset was collected using API. The process of identifying entities considers as a challenge if taking into account the nature of the components of the tweet.

The next challenge was to select the most relevant and crucial entities, this was determined by taking entities represent the person, location, and organizational entities. Ensures that these entities have a page on the site.

The next challenge was how to link to Wikipedia. The API is used for MediaWiki so that we can access the contents of the web page and extract the descriptive abstract of the Entity. Some of them have a web page in the wiki, but they do not have an abstract.mentioning that all he linking and extractions were performed online. These challenges might be avoided using different pooling techniques when earning the tweets stream, using different topic model algorithm. And finally linking to DBpedia might be used.

Referenes

[1] K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann, Twitter and Society. 2014.

[2] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, “Twitter part-of-speech tagging for all: Overcoming sparse and noisy data,” Proc. Recent Adv. Nat. Lang. Process., no. September, pp. 198–206, 2013.

[3] N. Gupta, S. Singh, and D. Roth, “Entity Linking via Joint Encoding of Types, Descriptions, and Context,” Emnlp, pp. 2671–2680, 2017.

[4] B. V. Barde and A. M. Bainwad, “An Overview of Topic Modeling Methods and Tools,” pp. 745–750, 2017.

[5] T. Shi, K. Kang, J. Choo, and C. K. Reddy, “Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local

Word-Context Correlations," Proc. 2018 World Wide Web Conf. World Wide Web - WWW '18, pp. 1105–1114, 2018.
[6] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic modeling over short texts," IEEE Trans. Knowl. Data Eng., vol. 26, no. 12, pp. 2928–2941, 2014.
[7] P. Wang, H. Zhang, B. Xu, C. Liu, H. Hao, and et al. Wang, Peng, "Short Text Feature Enrichment Using Link Analysis on Topic-Keyword Graph," Nat. Lang. Process. Chinese Comput., vol. 496, pp. 79–90, 2014.
[8] S. Hingmire, "WikiLDA : Towards More Effective Knowledge Acquisition in Topic Models using Wikipedia."
[9] W. Lukasiewicz, A. Services, and A. Paschke, "On the Move to Meaningful Internet Systems: OTM 2016 Workshops," vol. 10034, no. October 2016, 2017.
[10] L. P. Prieto, M. J. Rodríguez-Triana, M. Kusmin, and M. Laanpere, "Smart school multimodal dataset and challenges," CEUR Workshop Proc., vol. 1828, pp. 53–59, 2017.
[11] C. Lopez et al., "CAp 2017 challenge: Twitter Named Entity Recognition," 2017.
[12] R. Nugroho, D. Molla-Aliod, J. Yang, Y. Zhong, C. Paris, and S. Nepal, "Incorporating tweet relationships into topic derivation," Commun. Comput. Inf. Sci., vol. 593, pp. 177–190, 2016.
[13] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.
[14] T. Yano and M. Kang, "Taking advantage of Wikipedia in Natural Language Processing."
[15] A. Yıldırım, S. Üsküdarli, and A. Özgür, "Identifying topics in microblogs using wikipedia," PLoS One, vol. 11, no. 3, pp. 1–20, 2016.
[16] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia link structure and text mining for semantic relation extraction towards a huge scale global web ontology," CEUR Workshop Proc., vol. 334, pp. 59–73, 2008.
[17] G. Lansley and P. A. Longley, "The geography of Twitter topics in London," Comput. Environ. Urban Syst., vol. 58, pp. 85–96, 2016.