Tweets Classification, Hashtags Suggestion and Tweets Linking in Social Semantic Web

Rehab S. Ghaly
College of Computers and Information Technology
Arab Academy for Science and Technology and Maritime Transport
Cairo, Egypt

Emad Elabd
Faculty of Computers and Information
Menoufia University
Menoufia, Egypt

Mostafa Abdelazim Mostafa
College of Computers and Information Technology
Arab Academy for Science and Technology and Maritime Transport
Cairo, Egypt

***Abstract*—Nowadays social semantic Web has become one of the most important sources of data. The quality of data is crucial for data integration and annotation in social Linked Open Data (LOD). The main goal of tweets classification and linking is to integrate all tweets that are related semantically in one place on the Web. The data inequality problems exist because authors post tweets and hashtag them based on their own preference. The authors can post tweets in different languages (English- Arabic). Therefore, the hashtags cannot be the only trackback function of the tweet to link it with the related tweets. Thus, the goal of this paper is to develop an automated tweet classifier which is not limited by the language or the category of the tweet. The approach can create new suggested multilingual (English, Arabic) hashtags from tweet's content and comments "auto-tagging" in social linked open data (LOD), matching the tweets in the same topic and provides high accuracy and performance using blocking and indexing techniques. The proposed framework has substantial improvements in tweets classification and linking compared to state of the art framework of classifying and linking techniques with better classifying rate which reach 95% in precision and 97% in recall.**

Keywords—Social Semantic Web; Data Annotation; Tweet linking; Tweet classification; Hashtags suggestion

I. INTRODUCTION

Through past years, many researches have been done in social semantic Web [1]. Social semantic Web is considered a combination of two concepts (social networking and semantic Web). Social networking is a real time micro-blogging like twitter and Facebook where users always receive, send messages and share information. In the other hand Semantic Web (Linked Open Data) is creating typed links among data from multiple sources in the Web and construct global data space using Resource Description Framework schema (RDFs) [2]. Semantic Web is a set of many Ontology include a lot of RDF triples (subject, predicate and object). [2]. LOD links the different datasets which differentiate the relations between things to make the browsing much easier for users [2] [3].

Social networks are considered the fastest growing Web 2.0 services [4] [5]. The users can post short messages on social semantic Web which connected by trackback function [6]. In social semantic Web the users are using the hashtag (#) as a trackback function. Hashtags is a keyword without spaces between and proceeded by (#) sign to annotate data in social semantic Web [7] [8] [9].

There are three categories of social tagging (folksonomy, social tagging systems and tagging behavior of users) [10] [11]. The prediction of social hashtags belongs to the second category which used to enhance the performance and accuracy of the system. The researches in this area can be classified into three categories (1) How to determine subject from content of the text [12]. (2) How to predict new keywords in the same topic based on the existed hashtags [13] [14]. (3) How to link resources using hashtags [15] [16].

The quality of data which stock piling in LOD can have considerable cost effect on a system that uses the information to manage business [17]. Data integration is needed in social semantic Web to enhance the quality of data. Nowadays, there are millions of users shared a lot of posts daily in social semantic Web and most of them contain hashtags [18]. The main goal of hashtag in social semantic Web is the integration of different data which represent the same real-world objects. The posts may be written in bilingual languages like (English, Arabic). Arabic is the language of millions of people in the Middle East and North African countries [19] [20]. Arabic language is using in published data sources. That is why we need to integrate data in different languages in social semantic Web.

There are a lot of published datasets in LOD which present the social networking information as example (FOAF Profiles). Social networking datasets present all available data that can be found about the person and his social life like (posts, comments…etc.). Hashtags can be used as a matching criteria for those same semantically posts across different data sources. Since the same semantically post can be presented in different data sources in different languages (English- Arabic).

Posts written in English cannot directly be matched to posts written in Arabic due to different language script and morphology like sematic and linguistics problems. We propose in this paper an automated tweet classifier which classifies bilingual tweets. The approach can create new multilingual (English, Arabic) hashtags from tweet content in social LOD, linking the related tweets semantically and provides high accuracy using blocking/indexing technique [21].

The remainder of the paper is ordered as following: Section II explains the question in tweets classification and hashtag suggestion. Section III presents an overview about the work done in social semantic field and tweets classification. Section

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

IV describes the proposed framework and how it can help in classifying multilingual (English- Arabic) tweets and improving the suggestions of hashtags. Section V shows the result using the proposed framework and also in comparison to latest frameworks available. Section VI concludes the paper and shows the future work.

II. Problem Definition

Social media content is strongly multilingual. A lot of users shared posts daily and less than 50% of tweets are in English and the rest of percentage is in other languages. In the other side, there are only 8% of tweets contain a hashtags [22]. Hashtag is very important to link the tweets that related to each other in one place. In social semantic Web, authors post tweets and hashtag them based on their own personal preference. Therefore the hashtags in tweet's content can't be the only source used to classify the tweet.

Each author can hashtag the same tweet by using different hashtags and languages and it causes a problem when trying to link the tweets in different languages that related to each other in meaning. Posts written in English cannot directly be matched to posts written in Arabic due to different language script and morphology like sematic and linguistics problems. The Web contains millions of tweets which can be a problem in comparison process since the generated number of pairs will be huge.

Therefore, the contributions of this paper are important for many purposes. Firstly, it proposes an automated technique that classifies tweets using correct domains in social semantic Web. Secondly, the framework has the ability to suggest semantically related hashtags to describe the main objective of the tweet. Thirdly, the framework has the ability to link tweets from different domain and languages. Finally, the framework will show a high performance due to the using of blocking/indexing technique [21].

III. Related Work

Social semantic Web is one of the fastest growing Web 3.0 services. Users can post short messages and add hashtag which enables users to collect all related tweets in one group [10]. Hashtag prediction is a problem which appeared before in databases integration and a lot of works have been done in this area [6] [18]. There are a lot of techniques used to match and classify tweets and suggest hashtags for tweets. The traditional database community was discussed this problem before [18] [23] [24] [25]. The techniques used to classify the tweets as match or no match like:

- **Edit distances**: This technique compares two tweets to prove the dissimilarity by numeration the operation needs to convert one tweet to another [26].
- **Levenstein distance**: This technique calculates the number of replacements needed to convert one string to another [26].
- **N-gram**: This technique splits the tweet into substrings of length n. Then count how many n-grams are common in both tweets [27].
- **Jaccard distance**: In this technique, the number of words in pair of tweets which are having common characters are split by the whole number of the single words in every tweet [28].

TF-IDF and Soft TF-IDF are two of the well-known techniques that are used to measure the importance of each word in the tweet. TF-IDF technique is used to reflect the importance of each word due to the tweet [29]. Soft TF-IDF technique uses Jaro Winkler technique that works on tweets equality rather than exact match [30]. The Naïve bayes technique is used to determine the similarity between two tweets in the classification process [31].

Nadeau and Sekine reviewed a research for a diversity of languages and discussed a diversity of features, such as: (a) Contextual features, (b) Character-level features, (c) Part-of-speech (POS) and morphological features, (d) Gazetteers: extracting a large gazetteer from Wikipedia category names and redirects [32]. A lot of works have been done in bilingual (English- Arabic) hashtag prediction on social media but the problem still exists because of the lack of gazetteers [33] [34].

Li et al. discussed an unsupervised system that extracts keywords from the posts of social media Websites like Twitter posts and put them in a dictionary of possible hashtags. They discussed the features in the keywords then measure the most importance and close of each feature in the algorithm of the keyword selection [35]. Phan et al. presented a framework for classifying sparse and short text Web with hidden topics from huge collections of data [36]. Because of the shortness of tweets, they mentioned that a future work will be applied to develop other techniques to minimize the sparseness of data [36].

Esparza et al. categories the tweets into five classifications based on the content of tweet: books, games, movies, music and apps [23]. They classify each post manually. The classifier was trained to predict the categories of new tweets [23]. Li and Wu suggested hashtags from related tweets by using the similarity information of WordNet using a Euclidean distance metric as well [10]. Mazzia and Juett used the probability distributions to recommend hashtags. They chose naïve Bayes classifier to determine the hashtag is relevant in the individual tweet [24].

For Arabic, Darwish tested a system that was trained on news data on Arabic tweets. He reported results that were far lower than those for news [25] [33]. A lot of work has been done by Benajiba et al. on Arabic Named Entity Recognition. They used a gazetteer, a stop word list, current, previous, and next words, base phrase chunking, adjectives indicating nationality, cross-language capitalization and POS tagging. For evaluation, they created a dataset called ANERCORP [37]. The recent work done by Darwish used multilingual (English-Arabic) features from English to enhance Arabic NER. Most of work has been done on Arabic NER were focused on news [34].

Kareem Darwish and Wei Gao introduced simple effective language-independent approach to enhance named entity recognition (NER) on social media Websites.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

They used a two-pass semi-supervised technique, adaptation of domain, and big gazetteers. They compared the relative effectiveness of the approaches using (English-Arabic). The disadvantages of this approach were that it's needed predefined categories (domains) and it used trigger words to classify the tweets which are not sufficient enough [34].

To the best of our knowledge and experiments with the existing frameworks and techniques, they neither support classifying and linking multilingual tweets nor multilingual hashtag prediction in social semantic Web.

IV. Tweets Classification, Hashtags Suggestion and Tweets Linking Framework

The proposed framework is a Web-based hashtags suggestion and tweets linking. Hashtags suggestion and tweets linking framework is purposed and implemented to cope the disconnection between related tweets over the Linked Open Data (LOD). The architecture of the proposed framework consists of Datasets selection, Tweets Triples Listing, Data cleaning and standardizing, removing stop words, predicting tweet domain and finally suggesting hashtags as shown the framework architecture in Fig 1.

Tweets over LOD represented in different languages (Arabic-English) are selected as an input for the framework. Tweets Triples that contain tweets content and hashtags are used in the framework. Data cleansing and standardization is required for insuring the quality of data.

The proposed framework provides hashtags suggestion for existing tweets over the LOD, stop words get removed from tweets to keep only the keywords in each tweet. Keywords in the content of tweets are used to find the main domain and to categorize the tweet, keywords are translated and keywords-

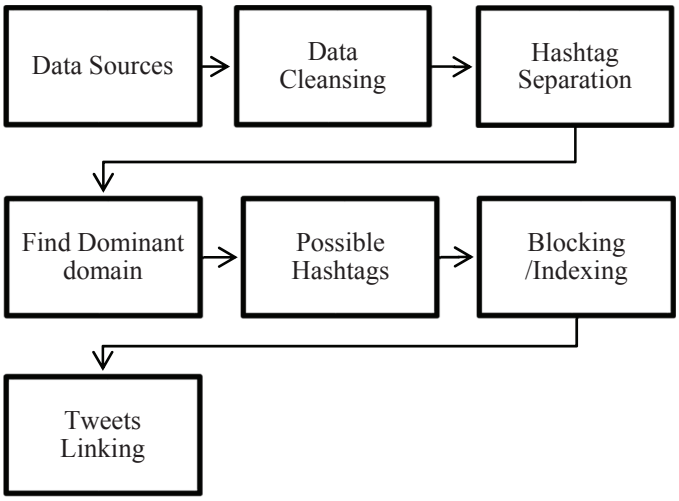

Fig. 1. Cross-language Name Matching Architecture

synonyms are found to collect and search for all available domains for the tweet then main domain is selected. Suggested hashtags are a collection of selected keywords synonyms and translations that are selected based on main domain. Suggested hashtags and main domain for each tweet gets used as an input for the linking process.

The following sub-sections discuss the framework in details.

A. Data sources

Tweets datasets were selected from FOAF:Tweets and yago [38] datasets. Two datasets were used as an input for the framework, first in English and second in Arabic. Tweets datasets contain data like (Tweet content, hashtags). Datasets are converted into RDF Graphs so tweets data is represented in a form of triples as shown in fig 2.

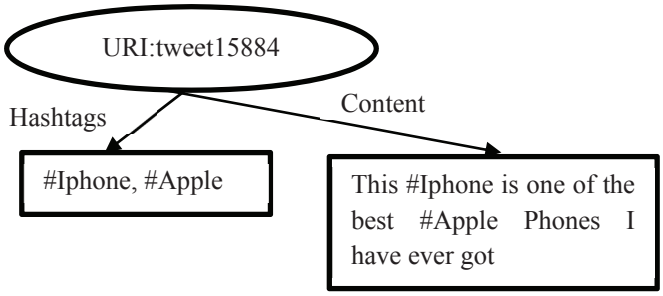

Fig. 2. Example on RDF Triples

B. Data Cleaning and Standardizing

The proposed framework uses triples containing tweets as an input for cleaning process. Data cleaning is the process of removing all inconsistent or rubbish data like (null triple, single word tweet, "___"), example on inconsistent data is finding only numbers in tweet content property or finding characters which does not form any kind of word in any language.

In the datasets, tweets content contain stop words which does not affect the meaning of the tweet. Stop words can be verbs or nouns [39]. An examples for these stop words in English are like: (with – and – I – is – her – my – won't – will – most), these stop words can be represented in different languages like Arabic as {(هو 'Howa', he), (هم 'Hom', they), (يكون 'Yakoon', is)… } where each item for the sake of explanation is represented in the form (Arabic word 'pronunciation in Arabic', Meaning in English) [39].

Stop words are removed from tweet content in the selected triples so the remaining content string in the tweet is the words that make difference in the meaning of the tweet content and can be used to propose hashtags. Finally the tweets get cleansed from inconsistent data.

C. Hashtags separation

Tweets hashtags are the tags used to identify the tweet topic and can be considered as keywords for the tweet. Hashtags are inserted into the tweet content in a form of a word starting with a hashtag sign "#". Hashtags are included inside the content of the tweet. Organic hashtags are considered as the initial sign of identifying the tweet category and domain which it follows and can be linked to.

Tweets contents get searched looking for hashtags. Hashtags get excluded from the tweet content and be used as initial possible hashtags suggestion. Fig. 3 shows the process of excluding hashtags from tweets content.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

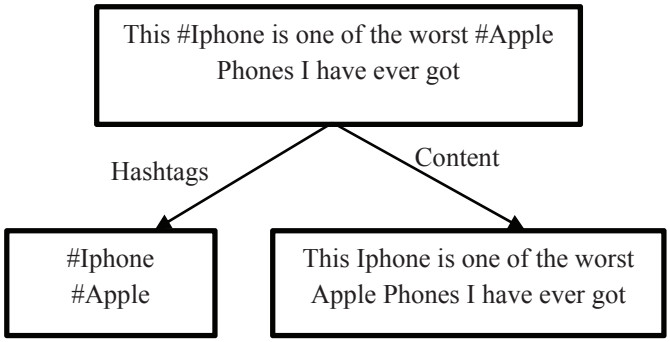

Fig. 3. Example of hashtags excluding process

Tweet content is parsed looking for hashtags inserted in the content. Once a hashtag sign is found, following text is saved as initial possible hashtag for the tweet. Hashtags can be categorized under different domains. Classification process is used to find the dominate domain for the tweet.

D. Dominant domain

Tweets get posted to comment on a specific matter. This matter can be classified under a specific domain or subject. Each tweet can be analyzed to fall under more than one domain, but mainly it should has a dominant domain. Organic hashtags can be miss-leading, an example for miss-leading hashtags is "#disappointed" while tweet can be talking about personal opinion on a new car or mobile device, since this hashtag describe an emotion instead or mentioning the device. People post tweets and hashtag them based on their own personal preference, therefore organic hashtags in tweet's content can't be the only source used to determine the dominant domain of the tweet.

Tweet content gets analyzed to find more possible hashtags that can be produced from the tweet content. Tweet content gets refined by removing stop words, links and symbols. Then, the remaining words are used as hashtags. Each hashtag can fall under one or more domains, so a classification and ranking process start to search for the dominant domain based on the occurrence of a specific domain more than others based on classifying all words in the tweet content under their domains from a lookup lexicon on Linked Open Data (LOD). Dominant domain is added to the tweet hashtags as an addition hashtag for this tweet. The process proceeded as following:

Get: *tweet content*
Create: *array of word from tweet content*
For: *words in array*
Get possible domains
Create: *list of possible domains*
For: *words in array*
IF: *domain is found add 1 to occurrence of domain*
Sort: *domains get sorted and Top 1 domain is used as dominant domain*

E. Possible Hashtags suggesion

Hashtags can be so useful for the user to determine the subject for the tweet, it can also be used as a searching criteria. Adding hashtags to a tweet, based on its subject, can help in classifying the tweet under correct domain or category.

Each word in the tweet content can have more than one synonym, and each synonym can have more than one meaning if it was used in different order of words. In addition to that tweets can be found in different languages like (English – Arabic) and both tweets in different languages can have the same meaning or at least fall under the same dominant domain or category. Table 1 shows sample of words that can have more than one meaning and translations.

TABLE I. WORD TRANSLATION, SYNONYM AND DOMAINS

Word	Translation /synonym	Pronunciation in English	Possible dominant domains
apple	تفاحة	Tofaha	fruit
apple	التفاح	Al tofah	fruit
apple	شجرة التفاح	Shagaret al tofah	fruit
apple	شركة أبل	Sherket apple	company, device
apple	ابل	apple	company, device
apple	apple company	-	company, device
apple	applecompany	-	company, device

Tweet content is used to suggest new hashtags to tweets, words in the content get translated to the preferred language (in this case to Arabic or English) and based on the dominant domain of the tweet synonyms and translations for each word get selected. The original words, selected translation and synonyms are added to the tweet as suggested hashtags. The process proceeded as following:

Get: *tweet content and dominant domain*
Create: *array of words from tweet content*
For: *words in array*
Find: *translations in different languages (English – Arabic)*
For: *each translation*
Find: *translation domain*
IF: *translation domain equal to dominant domain then add translation as suggested hashtag*

F. Indexing/Blocking

Since millions of tweets are found in Linked Open Data datasets, when comparing two dataset for finding linked tweets, each tweet from dataset A need to be compared with every tweet from dataset B, resulting a combination of comparative pairs |A x B| [40]. Tweets in the same dataset also needs to be linked, this means each tweet should be compared with each other tweet in the dataset which result N^2 number of compared pairs where N is the number of tweets in the dataset.

Traditional blocking technique was used to index the tweets into sorted blocks which reduce the number of compared pairs so that only possible matched tweets are grouped into blocks [41]. Dominant domain was used as a Blocking Key Value (BKV) to split tweets in each dataset into blocks.

After finding the dominant domain and adding suggested hashtags [21]. Each tweet is added to the index using its Unified Resource Identifier (URI) and dominant domain as BKV. Index gets generated for the datasets and saved for linking step. Table 2 and fig 4 shows example of indexing process using dominant domain as BKV.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

TABLE II. EXAMPLE OF TWEETS INDEXED USING DOMINANT DOMAIN AS BKV

Identifiers	Tweets URI	BKV (Dominant domain)
I1	:D1T1	Mobile
I2	:D1T2	Mobile
I3	:D1T3	Fruit
I4	:D1T4	Car
I5	:D1T5	Car

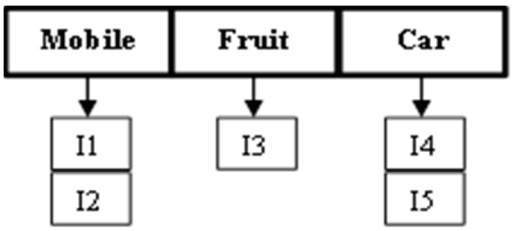

Fig. 4. inverted index as used for traditional blocking

G. Tweets linking

Adding dominant domain and suggested hashtags is critical tweets linking since it helps in determining the real category for each tweet. Blocks of tweets are created based on their dominant domain, each block can have tweets in different languages as long as they have same dominant domain. Each pair of tweets are compared based on organic hashtags and suggested hashtags. If hashtags in one tweet match hashtags in another, SameAs link gets created between them referring that both hashtag have the same meaning of discuss the same subject. Table 3 and table 4 show two tweets from different languages linked based on their suggested hashtags and dominant domain.

TABLE III. EXAMPLE 1 FOR ENGLISH TWEET

Tweet content	Organic hashtags	Dominant domain	Suggested hashtags
This IPhone is one of the worst Apple Phones I have ever got #disappointed	#disappointed	Mobile	#IPhone, #Apple,# ابل#,ايفون

TABLE IV. EXAMPLE 2 FOR ARABIC TWEET

Tweet content	Organic hashtags	Dominant domain	Suggested hashtags
هذا الاصدار من هاتف الايفون هو #الاسوأ اصدار من شركة ابل علي الاطلاق	#الاسوأ	Mobile	#Iphone, #Apple,#ايفون ابل#,

V. EXPERIMENT AND RESULTS

In order to test our approach, a Simple Effective Microblog Named Entity Recognition (SEMNER) framework was chosen for comparing the performance on the proposed framework [34]. English and Arabic datasets were selected to perform number of experiments to check the performance of the framework. A server with Xeon processor and 16G Ram was used for running the experiments, this server has windows server 2008 R2 as operating system, Visual Studio 2013 (C# language) and DotNetRDF were installed as programming tools.

Experiment 1: *Sample of 1000 tweet (without blocking and indexing).*

In this experiment we used English and Arabic datasets containing 1000 tweets in three topics: Locations, Persons and Organizations. Table 5 below shows the results of precision and recall for this experiment.

TABLE V. EXPERIMENT 1 RESULTS

Quality metric	Proposed Framework
No. of Entities	1000
True Positives (TP)	645
True Negatives (TN)	0
False Positives (FP)	232
False Negative (FN)	123
Precision (TP/(TP+FP))	74 %
Recall (TP/(TP+FN))	84%
F-Measure	78%
Accuracy	65%

Experiment 1 results 74% in precision and 84% in recall. After investigating these results, we found that some English and Arabic words were not found in the used dictionary, since this dictionary was outdated and does not contain slang words and expressions. Online lexicon was used which has huge number or words in different languages (English – Arabic), also it contain the latest slang words and expressions with a specification of its subject or domain.

Experiment 2*: Sample of 1000 tweets with the use of online lexicon (without blocking and indexing)*

In experiment 2, same sample of tweets was used in experiment 1 was used in experiment 2. Table 6 below shows the results of precision and recall for this experiment after replacing the lookup dictionary with an online updated lexicon.

TABLE VI. EXPERIMENT 2 RESULTS

Quality metric	Proposed Framework
No. of Entities	1000
True Positives (TP)	965
True Negatives (TN)	0
False Positives (FP)	21
False Negative (FN)	14
Precision (TP/(TP+FP))	98 %
Recall (TP/(TP+FN))	99%
F-Measure	98%
Accuracy	96%

Experiment 2 results 98% in precision and 99% in recall which is a dramatic increase in the quality and the accuracy of the framework. After investigating the results, we found that the performance of the framework decrease when we use large number of tweets due to the large number of generated pairs of tweets to be compared, this means that the time needed to perform the linking process is inversely proportional with the number of tweets.

Tweets were indexed based on their hashtags and divided into blocks to overcome the performance problem, since creating separate blocks of tweets minimize the number of pairs of tweets to be compared this lead to better performance, and the framework was not affected by the number of tweets needed to be linked.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

Experiment 3: *Comparing framework performance after using blocking and indexing*

In experiment 3, same sample of tweets used in experiment 2 was used in experiment 3. Table 7 below shows the number of pairs generated from the proposed framework and after using blocking and indexing.

TABLE VII. EXPERIMENT 3 RESULTS

	Proposed framework without blocking and indexing	**Proposed framework with blocking and indexing**
No. of tweets	1,000	1,000
No. of generated pairs	1,000,000 pairs	300,000 avg pairs in each block
Average computation time (seconds)	414 seconds	127 seconds

Experiment 3 results shows that both frameworks have the same complexity which is equal to O (n). Using blocking and indexing reduced the number of pairs generated and the time consumed to compute the linking process [21]. The time and number of pairs generated after adding the blocking/indexing technique is equal to 0.3 of the previous version of the proposed framework.

Experiment 4: *Comparison between SEMNER and the proposed framework.*

In this comparison, 10000 thousand tweets in different languages (English and Arabic) were extracted from FOAF: Tweets [10] and yago [38] datasets, those tweets were in three topics: Locations, Persons and organizations. Comparing the proposed framework with the latest similar framework for tweets classifying which is SEMNER [34], we found an improvement in the results as shown in Table 8 and Table 9.

TABLE VIII. COMPARISON BETWEEN RESULTS BETWEEN SEMNER AND PROPOSED FRAMEWORK

Quality metric	**SEMNER**	**Proposed Framework**
No. of Entities	10,000	10,000
True Positives (TP)	6987	9254
True Negatives (TN)	0	0
False Positives (FP)	384	481
False Negative (FN)	2629	265
Precision (TP/(TP+FP))	95%	95%
Recall (TP/(TP+FN))	73%	97%
F-Measure	82%	96%
Accuracy	70%	93%

TABLE IX. PERFORMANCE COMPARISON RESULTS BETWEEN SEMNER AND PROPOSED FRAMEWORK IN TERM OF COMPUTATION TIME

	SEMNER	**Proposed framework with blocking and indexing**
N. of tweets	10,000	10,000
No. of pairs	100,000,000 pairs	30,000,000 avg pairs in each block
Average computation time (seconds)	516 seconds	182 seconds

Comparing the proposed framework results with SEMNER framework which is the state of the art in tweets classifying and linking, we found that we have the advantage of higher accuracy rate which exceed 95% in precision and 97% in recall. Using blocking/indexing technique affected the number of pairs to be compared and the computation time, both factors decreased which lead to better performance. The proposed framework is fully automated and can work on any topic or domain in different languages (English – Arabic) of tweets which is an advantage over the SEMNER that can only work on three topics. The proposed framework is using blocking and indexing techniques which keep the performance independent from the number of tweets used for classification of linking process. Finally the proposed framework is using an online lexicon as a lookup for words which keep it up to date.

VI. CONCLUSION AND FUTURE WORK

The data in social semantic Web is growing rapidly. Therefore the data integration plays an important part in social LOD. In this paper, a framework for classifying and linking multilingual tweets and multilingual hashtags suggestion in social semantic Web is proposed with enhancing the accuracy and performance using blocking and indexing techniques. The proposed framework helped in classifying tweets and expected it can be applied in tweets which written in other languages over social LOD. The proposed framework has compared to state of the art framework of classifying and linking techniques with better classifying rate which reach 95% in precision and 97% in recall. In the future work, we will work on linking tweets with published articles under the same topics which takes in consideration the classification methods which will increase the precision and recall rate and give much better results in cross language classifying and linking.

REFERENCES

[1] Sami Mäkeläinen, "Social Semantic Web," in Paper for" Tiedonhallinta Semanttisessa Webissä"-seminar, University of Helsinki, 2005.

[2] Tim, Christian Bizer, and Tom Heath Berners-Lee, "Linked Data - The Story So Far," International Journal on Semantic Web and Information Systems, pp. 1-22, 2009.

[3] V., and S. Chenthur Pandian Subramaniyaswamy, "A complete survey of duplicate record detection using data mining techniques," Information Technology Journal 11, 2012.

[4] Scott, and Sarita Yardi Golder, "Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality," in Social Computing (SocialCom), 2010 IEEE Second International Conference, 2010, pp. 88-95.

[5] Danah, Scott Golder, and Gilad Lotan Boyd, "Conversational Aspects of Retweeting on Twitter," in System Sciences (HICSS), 2010 43rd Hawaii International Conference, 2010, pp. 1-10.

[6] Tim Berners-Lee, "The Social Semantic Web," in W3C semantic Web, 2013.

[7] Gabor, and Bernardo A. Huberman Szabo, "Predicting the Popularity of Online Content," in Communications of the ACM 53.8, 2010, pp. 80-88.

[8] Stanley, and Katherine Faust Wasserman, "Social Network Analysis: Methods and Applications," in Cambridge University Press, vol. 8, 1994.

[9] Kalina, Hamish Cunningham Bontcheva, Semantic Annotations and Retrieval: Manual, Semiautomatic, and Automatic Generation.: In Handbook of semantic web technologies, Springer Berlin Heidelberg, 2011.

[10] Tianxi, Yu Wu, and Yu Zhang Li, "Twitter Hash Tag Prediction Algorithm," in ICOMP'11-The 2011 International Conference on Internet Computing, 2011.

[11] Jennifer Trant, "Studying social tagging and folksonomy: A review and framework," Journal of Digital Information 10, pp. 1-44, 2009.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE

[12] Hendri, and Klaus Obermayer Murfi, "A two-level learning hierarchy of concept based keyword extraction for tag recommendations," ECML PKDD Discovery Challenge, pp. 201–214, 2009.

[13] Ralf, Peter Fankhauser, and Wolfgang Nejdl Krestel, "Latent Dirichlet Allocation for Tag Recommendation," in Proceedings of the third ACM conference on Recommender systems, 2009, pp. 61–68.

[14] Ralf, Peter Fankhauser Krestel, "Tag Recommendation Using Probabilistic Topic Models," in ECML PKDD Discovery Challenge, 2009, pp. 131–141.

[15] Yu-Ta, Shoou-I. Yu, Tsung-Chieh Chang, and Jane Yung-jen Hsu Lu, "A Content- Based Method to Enhance Tag Recommendation," IJCAI, vol. 9, pp. 2064-2069, 2009.

[16] Shankara B., Huan Liu Subramanya, "Socialtagger-collaborative tagging for blogs in the long tail," in Proceedings of the 2008 ACM workshop on Search in social media, 2008, pp. 19-26.

[17] ens, and Felix Naumann Bleiholder, "Data Fusion," in ACM Computing Surveys. (CSUR) 41, no. 1, 2008.

[18] Su Mon, Ee-Peng Lim, and Feida Zhu Kywe, "A survey of recommender systems in twitter," in Social Informatics, 2012, pp. 420-433.

[19] Tim Berners-Lee, "Semantic Web Road Map," in W3C Draft http://www. w3. org/DesignIssues/Semantic, 1998.

[20] Layan M. Bin, and Hend S. Al-Khalifa Saleh, "AraTation: An Arabic Semantic Annotation Tool," in Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, 2009, pp. 447-451.

[21] Peter. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," in Knowledge and Data Engineering, IEEE Transactions on 24, 2012, pp. 1537-1555.

[22] Robin Wauters. (2010, feb) twitter language. [Online]. http://techcrunch.com/2010/02/24/twitter-languages/

[23] Sandra, Michael P. O'Mahony, and Barry Smyth. Garcia Esparza, "Towards Tagging and Categorisation for Micro-blogs," in the 21st National Conference on Artificial Intelligence and Cognitive Science (AICS 2010), Galway, Ireland, 2010.

[24] James Juett Allie Mazzia, "Suggesting Hashtags on Twitter," EECS 545 Project, Winter Term, University of Michigan, 2011.

[25] Ahmed, and Kareem Darwish Abdul-Hamid, "Simplified Feature Set for Arabic Named Entity Recognition," in Proceedings of the 2010 Named Entities Workshop, Association for Computational Linguistics, 2010, pp. 110–115.

[26] Andrew T., Sherri L. Condon, and Christopher M. Ackerman Freeman, "Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance algorithm," in Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp. 471-478.

[27] Peter F., Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Brown, "Class-Based n-gram Models of natural Language," in Computational linguistics 18, no. 4, 1992, pp. 467-479.

[28] Seung-Seok, Sung-Hyuk Cha, and Charles C. Tappert Choi, "A Survey of Binary Similarity and Distance Measures," Journal of Systemics, Cybernetics and Informatics 8, vol. 8, pp. 43-48, 2010.

[29] Ho Chung, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok Wu, "Interpreting TF-IDF term weights as making relevance decisions," in ACM Transactions on Information Systems, 2008.

[30] Michelle, and Pascal Hitzler Cheatham, String Similarity Metrics for Ontology Alignment. Berlin Heidelberg: The Semantic Web ISWC, Springer International Publishing AG, 2013.

[31] David D Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in Machine learning: ECML-98, 1998, pp. 4-15.

[32] S Sekine D Nadeau, A Survey of Named Entity Recognition and Classification. New York, USA: John Benjamins Publishing Co, Lingvisticae Investigationes 30, April 2007.

[33] Kareem Darwish, "Named Entity Recognition using Cross-lingual Resources: Arabic as an Example," in ACL (1), 2013, pp. 1558-1567.

[34] Kareem, and Wei Gao Darwish, "Simple Effective Microblog Named Entity Recognition: Arabic as an Example," in The International Conference on Language Resources and Evaluation, 2014.

[35] Zhenhui, Ding Zhou, Yun-Fang Juan, and Jiawei Han Li, "Keyword extraction for social snippets," in Proceedings of the 19th international conference on World wide web, 2010, pp. 1143-1144.

[36] Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi Phan, "Learning to classify short and sparse text \& web with hidden topics from large-scale data collections," in Proceedings of the 17th international conference on World Wide Web, New York, NY, USA, 2008, pp. 91-100.

[37] Yassine, Mona Diab, Paolo Rosso Benajiba, "Arabic named entity recognition using optimized feature sets.," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 284–293.

[38] datahub. (2013) [Online]. http://datahub.io/dataset/yago

[39] Ranks nl. (1998) [Online]. http://www.ranks.nl/stopwords

[40] Peter, and Karl Goiser Christen, "Quality and complexity measures for data linkage and deduplication.," in Quality Measures in Data Mining, pp. 127-151.

[41] Ivan P., and Alan B. Sunter. Fellegi, "A theory for record linkage," Journal of the American Statistical Association 64, pp. 1183-1210, 1969.

978-1-4673-8460-5/16/$31.00 ©2016 IEEE