Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 335–344

Procedia Computer Science

www.elsevier.com/locate/procedia

22nd International Conference on Knowledge-Based and
Intelligent Information & Engineering Systems

A Semantic Approach for Tweet Categorization

Ben Ltaifa Ibtihel[a,*], Hlaoua Lobna[b], Ben Jemaa Maher[a]

[a] *ReDCAD Laboratory, ENIS Sokra km 3.5, University of Sfax, Tunisia*
[b] *MARS Research Laboratory, LR 17ES05, University of Sousse, Tunisia*

Abstract

The explosion of social media and microblogging services has gradually increased the microblogging data and particularly tweets data. In microblogging services such as Twitter, the users may become overwhelmed by the rise of data. Although, Twitter allows people to micro-blog about a broad range of topics in real time, it is often hard to understand what these tweets are about. In this work, we study the problem of Tweet Categorization (TC), which aims to automatically classify tweets based on their topic. The accurate TC, however, is a challenging task within the 140-character limit imposed by Twitter. The majority of TC approaches use lexical features such as Bag of Words (BoW) and Bag of Entities (BoE) extracted from a Tweet content. In this paper, we propose a semantic approach of improving the accuracy of TC based on feature expansion from external Knowledge Bases (KBs) and the use of eXtended WordNet Domain as a classifier. In particular, we propose a deep enrichment strategy to extend tweets with additional features by exploiting the concepts present in the semantic graph structures of the KBs. Then, our supervised categorization relies only on the ontological knowledge and classifier training is not required. Empirical results indicate that this enriched representation of text items can substantially improve the TC performance.

Keywords: Tweet Categorization (TC); semantic approach; deep enrichment; Knowledge Bases (KBs); eXtended WordNet Domain

1. Introduction

Twitter is a major microblogging service providers. It has become the first and quickest source of relevant information, which allows people to micro-blog about a broad range of topics [26]. This social networking application lets the users present any information with only a few words, optionally followed by a link to a more detailed source of information. In microblogging services such as Twitter, the users may become overwhelmed by the raw data and the high topical diversity. Indeed, users are usually only interested in certain topic categories, such as politics, sports or entertainments. Therefore, TC can greatly simplify browsing large collections of tweets by reorganizing them into a smaller number of manageable categories. In this work, we study the problem of TC aiming to automatically classify

[*] Corresponding author. Tel.: +216 53 632 258 ; fax: +216 74 666 578
E-mail address: ibtihel.beltaifa@gmail.com

10.1016/j.procs.2018.07.267

tweets according to their topics. However, TC task poses different challenges including: abbreviations, misspelled, colloquial expressions and short form for words which make it hard to understand and extract meaningful semantics from tweets. Besides, the BoW representation used in traditional TC methods is often unsatisfactory as it ignores relationships between important terms that do not cooccur literally. In order to deal with this problem, many approaches integrate core ontologies as background knowledge into the process of tweet classification. In this paper, we propose a novel semantic approach to do feature generation for TC using eXtended WordNet Domains[1](XWND) as a classifier. This work explores a deep enrichment strategy to improve the performance of TC by extending the tweets with different semantic features extracted from external KBs such as DBpedia. The main goal of this work is to show that the semantic enrichment improves TC results. The rest of the paper is organized as follows: in Section 2 we present a review of relevant literature. Section 3 presents our motivation and contribution in this work. Section 4 details our semantic approach for TC. Section 5 presents and discusses the experimental evaluation of our semantic approach. A conclusion and future work discussion are given in section 6.

2. Related work

A number of methods have been proposed for tweet classification/categorization. Existing approach for tweet classification can be categorized into two main categories : traditionnal approaches and semantic approaches.

2.1. Traditionnal approaches

Traditionnal approches that use local metadata rely on the presence of words or syntactical features that represent tweet like URLs, hashtags,etc. These approches focus on words and cannot provide semantic information and relationships among them. Which case, a number of researchers in tweet classification task ([27], [17], [5], [2]) had been working on the machine learning methods based on training the classifier on a labeled text collection. These supervised learning methods include using Support Vector Machines (SVM), Neural Networks (NN), Naive Bayes (NB) and Random Forests over BoW approach to solve a lot of classification problems. Authors in [21] proposed a new tweet classification method that made use of tweet features like the URLs in the tweet, retweeted tweets and tweets from the most influential user of a trend topic. The performance of the proposed method, based on tweet metadata, increases the classification accuracy compared to the BoW approach used to represent the tweets. Another work was done by [11] for twitter trending topic classification. The authors used two different classification approaches: the BoW approach for text classification using the NB Multinomial classifier and the network-based classification approach which classify the given topic using the C5.0 decision tree learner. In [23], the authors used a small set of domain-specific features extracted from the author's profile and text for classifying tweets into predefined set of generic classes. Empirical results show that the authorship plays a crucial role in tweet classification. While in [24], authors studied how the Map-Reduce paradigm can be applied to NB classifier for tweet classification in order to handle large numbers of tweets. Moreover, authors of [18] presented an approach based on the comination of five classifiers for the classification Of Spanish Election Tweets (COSET). The best experimental results are obtained with lemmatized texts and the feature selection method that removes features with low variance. In contrast, authors in [19] studied the effect of adding POS (part-of-speech) tag unigrams and bigrams as new features on the performance of classifying disaster tweets using NB classifier. Experimental results have shown that the POS tags can improve the performance of the domain adaptation classifiers compared to the BoW representations only.

2.2. Semantic approaches

Semantic approaches, on the other hand, exploit the link structure of the knowledge Bases (KBs) (e.g. Freebase and Wikipedia). The majority of these approaches are based on feature expansion. For that, authors in [3] introduced a novel approach for topic classification of tweets using multiple linked knowledge sources (KSs) such as DBpedia, Yago and Freebase. They proposed a KS-based contextual enrichment of features. The experimental results using

[1] http://adimen.si.ehu.es/web/XWND

multiple linked KSs based on SVM classifier outperform the approach using a single KSs. Authors in [12] proposed to exploit paraphrases to improve tweet topic classification performance. They explored two approaches to generate paraphrases, WordNet and word embeddings. The experiment results shows that the word embedding approach outperforms the WordNet method. In [22], authors employed embedding words based on word2vec to address the problem of vocabulary mismatch for tweet topic classification. In this stady, they applyed feature expansion with three learning algorithms NB, SVM and Logistic Regression (Logit). Best result is obtained using Logit classifier. While in [20], authors presented a semantic abstraction to improve the generalization of tweet classification. They extracted features from Linked Open Data and included location and temporal mentions from tweet content. In [7], authors proposed an approach to classify Twitter messages by mappping each tweet to the most similar Wikipedia articles based on a semantic distance calculation. The results showed that latent semantic topic modelling approach based on lexical features extracted from tweet's content only performed better than the Latent Semantic Analysis (LSA) and the String Edit Distance (SED) methods. Autors in [4], however explored the idea of combining textual content with structural information provided by the relationship between tweets and users. Experimental results show that this combination proposal is effective in TC.

3. Motivation and contribution

Many tweet classification/categorization solutions presented in the previous section that only relate tweets using identical terminology and ignore conceptual similarity of terms as defined in terminological resources like WordNet [19]. Traditional supervised TC use machine learning methods ([27], [5], [2]) which uses annotated texts with given labels to learn a classifying model, or statistical approaches to perform the task. Such methods, including SVM, NB, decision trees, and Latent Semantic Analysis, are effective, but all of them require a set of pre-classified tweets to train the classifier. Moreover, processing Twitter messages is challenging because tweets are at most 140 characters long and they are too short to extract from them meaningful semantics on their own. Then, the normal BoW approach in short text classification would not perform well on tweets because of the shortness and fail to achieve high accuracy due to data insufficiency issue [11]. To tackle this limitation, many existing approaches exploited additional Knowledge Sources (KSs) to enrich tweet's content. These sources provide semantic features (e.g. BoE) derived from the tweet content only to enhance the lexical representation of text (e.g. BoW) ([25], [22], [4]). Furthermore, autors in [3] proposed a generic framework based on contextual enrichment of features using different linked KSs. This enrichment focus only on extending the named entities (e.g. persons, locations and organizations) present in tweets without taking into consideration that some tweets do not contain named entities and ignoring the specific meaning of each term that forms the tweet. To overcome these limitations, we propose a semantic approach based on supervised categorization which relies only on the ontological knowledge while classifier training is not required. In particular, we enrich the textual content of tweets by linking the tweets with additional features extracted from KBs.
Our contributions in this work are two-fold:

(1) An expansion technique to enrich and extend text tweets with additional semantic features: we propose a deep enrichment process for representing tweets using synsets from WordNet and concepts from DBpedia by exploiting the semantic features present in the graph structures of the KBs.

(2) A semantic approach based on deep enrichment strategy to automatically classify tweets into a specific category (domain) using XWND as a classifier. Our supervised categorization relies only on the ontological knowledge and classifier training is not required.

4. A semantic approach based on deep enrichment

This section describes the approach we propose to classify tweets into categories. Given a set of tweets as input, the main steps of our approach include a Pre-processing step to clean the input data, Named Entity (NE) Expansion based on NE Recognition and Linking, Word Sense Disambiguition and finally TC. The model of our semantic approach for TC presented in Fig. 1 highlights the following steps:

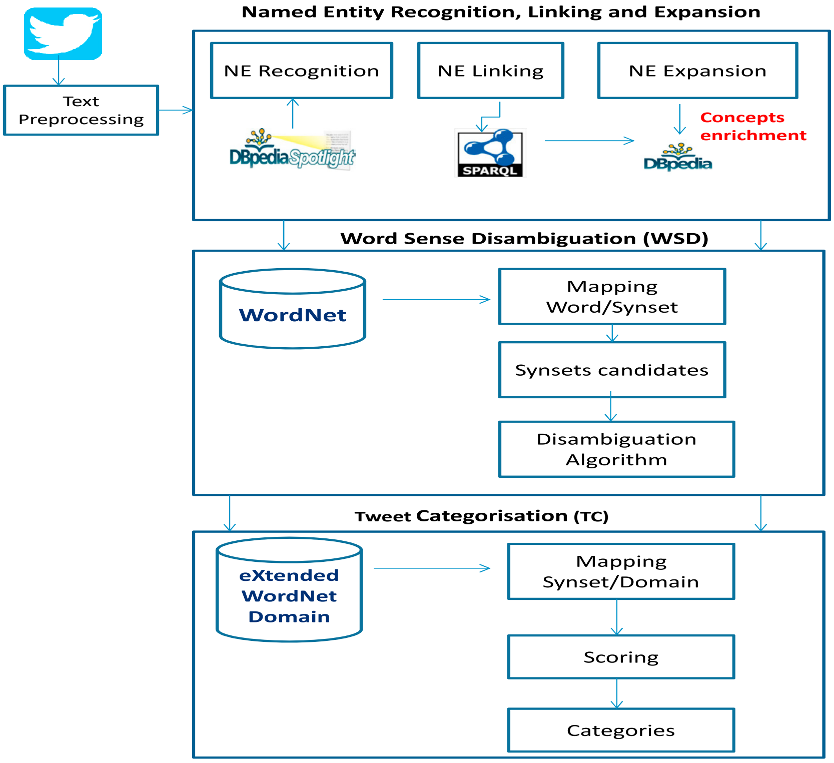

Fig. 1. Model of our semantic approach for TC

4.1. Pre-processing

Before tweets are analysed, they are pre-processed as follows:

- Eliminating non English tweets;
- Removing the word 'RT', user mentions '@' and URLs;
- Removing punctuation, stop words, numerical characters and special characters (emoticons);
- For hashtags the '#' symbol is removed and the tag is split according to a capital-letter rule used in [9]. For example, '#WhitneyHouston' becomes 'Whitney Houston';
- Spell check and automatic corrections. We use the Bing spell check API [2] for solving some misspellings;
- Each word of the tweets is lemmatized.

4.2. NE Recognition, Linking and Expansion

In this section, we describe the different steps 3-stage approach of the NE Recognition, Linking and Expansion phase. DBpedia is the source of knowledge for our work which extracts structured information from Wikipedia. We choose DBpedia because it is the most widely used general-purpose knowledge base in the Semantic Web and NLP community [6]. DBpedia allows users to query relationships and properties associated with Wikipedia concepts. In this paper we used SPARQL [3] to query DBpedia.

[2] https://www.programmableweb.com/api/bing-spell-check
[3] http://dbpedia.org/sparql

4.2.1. Named Entity Recognition

In text processing, Named Entity Recognition (NER) is referred to as task for designating specific keywords or tokens, phrases from within text to their identifying entity classes such as persons, locations and organizations [10]. In the named entity recognition, we focus on identifying named entities and their types in tweet texts. Here, the goal is the identification of surface forms (terms that are used to mention the entities in text)), and their labeling with one of several entity type labels. Note that an entity (such as New York) can be referred to multiple surface forms (e.g., New_York_State and New_York_City). For entity identification, we use DBpedia Spotlight[4], a system that links regular text to DBpedia. First, N-grams are spotted to refer to entities called surface forms, then the DBpedia lexicalization dataset is used to generate candidate entities for the spotted N-grams. There are three sources that we use to extract and identify surface forms (sources of mappings between surface forms and DBpedia resources): the titles of entity pages that offer chosen names for resources (An entity page is an article that contains information focused on one single entity, such as a person, a place), the titles of redirecting pages that offer alternative spelling, alias, etc, and the disambiguation pages that link a common term to many resources. For each mention, we have potentially many candidates, some of which have to be filtered out because they are not related to the context. In the disambiguation step, the best candidates are chosen.

4.2.2. Named Entity Linking (Disambiguation)

Once recognized (extracted and typed), entities are linked (disambiguated) according to a reference KB. Our disambiguation relies heavily on Wikipedia content. For disambuigating, DBpedia spotlight selects the correct candidate DBpedia Resource for a given surface form. The decision is made based on context surface form mention. Then, our system collects context for DBpedia ressources from Wikipedia. We use three types of context : wikipedia pages, definitions from disambiguation pages, paragraphs that link to ressources. For each mention, we have potentially many candidates, while some of them have to be filtered out because they are not related to the context. In order to filter, we use a ranking algorithm based on a similarity measure. DBpedia Spotlight relies on the so-called $TF * ICF$ (Term Frequency Inverse Candidate Frequency) score computed for each entity [15].

4.2.3. Named Entity Expansion: concept enrichment

In this step, we propose a semantic enrichment process for tweet representation. We use DBpedia not only to link to its entities but also to extract related categories (concepts) for feature expansion to enrich tweets. Then, after linking the tweet text to DBpedia entities (shallow enrichment), we also explore additional DBpedia concepts that are not directly observable in the tweet (deep enrichment). As an example, consider the following tweet: The latest Manchester City FC news and reports in BBC . With the shallow enrichment method, only the entities dbpedia: Manchester City FC and dbpedia: BBC would be added as extra features for classification. However, these features didn't characterize the main topic or context of the tweet. A solution to this problem is to use deep enrichment using concept expansion which provides rich semantic information from KB. The purpose of deep enrichment is to generate more features into the process of tweet classification. DBpedia provides an ontology of classes representing available knowledge about a vast number of concepts across Wikipedia pages (Wikipedia Categories). These concepts about different resources over Wikipedia are categorized under classes such as thing, agent, food, place, etc. (rdf : Types)[1]. Then, after selecting the correct candidate DBpedia Resources for a given entity mentions (surface form), we automatically extract from DBpedia all the related categories of the DBpedia resources. In other words, for every resource collected, we query the DBpedia SPARQL endpoint[5] to retrieve all the categories connected to the resources. A DBpedia resource is linked to its categories through the dc:subject property. In particular, we enrich the textual content of tweets by linking the NE to all related concepts in DBpedia ontology. In what follows, by a concept we mean an entity or category from DBpedia knowledge base for representing text tweets. For example, for the mention entity (Lionel_Messi), we replace the NE mentions with DBpedia concepts (dbc:FC_Barcelona_C_players, dbc:Argentine_footballers, dbc:FIFA_World_Cup_players).

[4] http://demo.dbpedia-spotlight.org/

[5] http://wiki.dbpedia.org/OnlineAccess#1.1 Public SPARQL Endpoint

4.3. Word Sense Disambiguation (WSD)

Although DBpedia provide rich semantics from background knowledge for extracting entities and concepts they cannot cover all existing concepts in the tweet. For example, in the case of this tweet 'Train your brain to prefer healthy foods', we cannot extract any concept with DBpedia. To resolve this problem, we propose using WordNet synsets and DBpedia concepts together for representing the tweet. As large lexical database, WordNet [6] groups english words into set of synonyms (synsets) that denote the same concept. Synsets are connected to one another through explicit semantic relations (e.g. hypernymy and hyponymy). For example, perennial is a kind of plant, perennial is hyponym of plant and plant is hypernym of perennial. Then, we used WordNet for extracting the synonyms of nouns and verbs (we eliminate the adjectives and adverbs in our process). This step needs a Word Sense Disambiguation (WSD) for determining which sense of a word is used when it appears in a particular context [14]. In this section, we formally describe the disambiguation algorithm based on Specification Marks (SM) Method combined with the Hypernym/Hyponym Heuristic [16]. The basic hypothesis of the SM method is that the higher the similarity between two words is, the larger amount of information is shared by two of its concepts. In this case, the information frequently shared by several concepts is indicated by the most specific concept that subsumes them in the IS-A taxonomy (hypernymy/hyponymy relations). A concept corresponds to a SM sculptured in the form of a tree. The input for this WSD algorithm is a group of words $W = \{w_1, w_2, ..., w_n\}$ in a context. Each word wi obtained from WordNet has an associated set of possible senses $Si = \{Si_1, Si_2, ..., Si_n\}$, and each sense has a set of concepts in the IS-A taxonomy. The WSD based on SM Method is given by two algorithms (see Algorithm 1 and Algorithm 2).

Algorithm 1 WordSenseDisambiguation (C : set of Concepts, $setSM$: set of Specification Marks

```
VAR
SM: Specification Mark
$S_nMax$: $S_nMax$ is the proper synset of $w_i$
Resolved: Boolean
begin
    $k \leftarrow 1$
    $Resolved \leftarrow False$
    while not (Resolved) do                          ▹ the word $w_i$ has not been disambiguated
        $S_nMax \leftarrow SenseWord(C, SetSM[k])$
        if $S_nMax$=NiL then
            k=k+1
        else
            $Resolved \leftarrow True$
        end if
    end while
return $S_nMax$;
    end
```

If the word w_i has not been disambiguated by the SM method, it will be left aside to be processed by a complementary Hypernym/Hyponym Heuristic. This heureustic is used to solve the ambiguity of the words that are not explicity related in WordNet taxonomy. Then, all the hypernyms/hyponyms of the ambiguous word w_i are extracted and we focus on synsets that have concepts (hypernyms/hyponyms) matching with some word from the context. Each synset in the hypernym/hyponym chain is weighed according to their depth in the subhierarchy. We choose the proper sense having the greatest weight calculated by this formula :

$$Weight(senses[k]) = argmax_{k=1..C.senses.length()} \sum_{i=1}^{depth(senses[k])} \frac{level(c_i)}{TotalLevel(senses[k])} \tag{1}$$

[6] https://wordnet.princeton.edu/

with c_i presents a concept in a branch of the tree in the IS-A taxonomy that leads to the direction k, $senses[k]$ denotes the synset identifier in WordNet taxonomy, $C.senses.length()$ denotes the size of the list of synsets for the concept c_i in WordNet taxonomy, $level(c_i)$ denotes the number of the level of the concept c_i in the hypernym/hyponym subhierarchy, $TotalLevel(senses[k])$ counts for each synset identifier the total levels of the concepts in the hypernym/hyponym subhierarchy and $depth(senses[k])$ denotes the depth of the synset in the hypernym/hyponym subhierarchy.

Algorithm 2 SenseWord (C : Concept, SM : Specification Mark)

```
VAR
SetOfConcepts: set of concepts in the IS-A taxonomy (hypernymy/hyponymy relations)
MyContext: is the context of $w_i$
begin

    $SetOfConcepts \leftarrow \emptyset$
    $MaxSetOfConcepts \leftarrow \emptyset$
    $S_nMax \leftarrow Nil$                    ▷ SnMax is the proper synset of $w_i$ to be returned
    for $i = 1$ to $C.senses.length()$ do
        $S_n \leftarrow C.senses(i)$    ▷ To each sense appearing in the stacks, the method associates the list of subsumed senses from the context
        if $FindDFSTree(SM, S_n) = True$ then  ▷ FindDFSTree is a function based on DFS (Depth First Search) algorithm that used for traversing tree or graph data structures
            $SetOfConcepts \leftarrow FindAllConcepts(SM)$
            $SetOfConcepts \leftarrow SetOfConcepts \cap MyContext$    ▷ assigning to each SM the number of context words subsumed
        else
            if $FindConcepts(C, SM) = True$ then
                $SetOfConcepts \leftarrow C$
            end if
        end if
        if $SetOfConcepts.size() > MaxSetOfConcepts.size()$ then
            $MaxSetOfConcepts \leftarrow SetOfConcepts$
            $S_nMax \leftarrow C.senses(i)$   ▷ looking for synsets that have concepts that match with some word from the context
        else
            if $SetOfConcepts.size() = MaxSetOfConcepts.size()$ then
                $S_nMax \leftarrow Nil$                ▷ the word $w_i$ has not been disambiguated
            end if
        end if
    end for
return $S_nMax$;
end
```

4.4. Tweet Categorization (TC)

It consists of assigning tweets to be categorized to specific categories as a supervised multi-class based on the domains already found in XWND. A keyword is linked to a WordNet synset and the domain label for this synset as well as the domain labels for the word in a tweet. In this work, we applied the domains extracted from XWND to classify tweets. We used a normalized version of XWND [8] which is an ongoing work aiming to automatically improve WordNet Domains [13]. XWND is based on well known PageRank (PR)(random-walk model initially developed for Web page ranking) algorithm in order to determine which synsets are more important for a domain. The XWND directory consists of 170 files. Each file contains a vector of 117,536 syntaxes sorted by weight, from the highest to the lowest. Thus, the most representative syntaxes for a given domain are in the first positions.

The Extended version of WordNet Domains [7] uses a more specific domain : By their nature, domains can be organized in hierarchies based on a relation of specificity. For instance, we can say that Football is a more specific domain than Sport, or that Business is more general than Finance.
The semantic approach for TC can be summarised as follows: Let

$$D = \{D_1, D_2, ..., D_n\}$$

be the set of XWND and let

$$C = \{c_1, c_2, ..., c_m\}$$

be the set of synonymous concepts aggregated in WordNet (synsets)
Now, let w_i be a word and let

$$Senses(w_i) = \{c_i | c_i \in C\}$$

with c_i being a sense for w_i (sense disambiguated within the given context in 4.3)
Let

$$T = \{w_1, w_2, ..., w_s\}$$

be the whole words in the tweet.
Then, for each sense $Senses(w_i)$ of word w_i in the tweet, we consider only the domain with the highest PR weight. XWND assigns a score to each pre-defined domain annotated $score(w_i, D_j)$.
The domain relevance function D^* for a word has the following denition:

$$D^* = argmax \sum_{\substack{\forall w_i \in T \\ \forall D_j \in XWND}} score(w_i, D_j) \tag{2}$$

Finally, the tweet is then assigned a label corresponding to the topic (domain).

5. Evaluation

This section describes the experiment that has been conducted in order to evaluate the the effectiveness of our semantic approach for TC. Our main goal here is to analyze and compare the different lexical and semantic feature representations used to classify tweets. The first aim of this experiment is to evaluate the accuracy of TC. We begin with presenting the data collection and after that, we discuss the results.

5.1. Data collection

We evaluated our sematic approach with a tweet collection which covers 1330 tweets collected via Twitter Search API and limited to a six specific topics. These tweets were manually labeled as belonging into one of the six classes namely Sports, Business, Technology, Entertainment, Politics and Education. Only English tweets are included in this study.

[7] http://wndomains.fbk.eu/

5.2. Experiment results and discussions

The purpose of these experiments is to determine the effectiveness of our semantic approach for TC using classification metrics. The performance of the TC is analysed using accuracy and sensitivity metrics. We used the accuracy as a metric for evaluating the overall effectiveness of our classifier and the sensitivity as a metric for measuring the effectiveness of our classifier to identify positive labels that are correctly classified.
Table1 shows the results of TC using lexical and semantic features.

Table 1. Results of TC using lexical and semantic features

Features		Accuracy	Sensitivity
Lexical Features (Baseline Features)	BoW	50.53%	50.61%
	BoE	15.01%	10.55%
Semantic Features	BoS	59.40%	58.97%
	BoE+synsets	50.27%	51.17%
	BoE+concepts	69.98%	71.76%
	BoE+concepts+synsets (our approach)	88.16%	87.62%

In order to highlight the contribution of our deep enrichment approach, we have compared the use of lexical and semantic features in TC process. The most commonly used features in text analysis are the Bow which represent a text as an unordered set of words. It assumes that words are independent from each other and also disregards their order of appearance. As the data consists of text strings, a BOW approach was used to represent the tweet as a set of words. From Table 1, a baseline approach for TC is the BOW approach used with XWND classifier. Then, for the two classification metrics, we observe that the use of BOW approach had poor performance on TC compared to our approach (as described below). The performance of BoW is improved by integrating external features generated from WordNet, DBpedia and their combination respectively. The extraction of Named Entity from DBpedia, used as features for classification (BoE) didn't show a high accuracy and sensitivity. In fact, it is much harder to detect entities in short Twitter messages. This is explained by the fact that some tweets do not contain named entities. Besides, the named entities sometimes direct us to the required domain.
However, the results obtained using the related DBpedia categories (concepts expansion) are better than using just BoE as only feature. This combination with semantic features (concepts) improves the results. The BoS (Bag Of Synsets) from Wordnet based on context WSD achieves much better performance than BoE. We conjecture that it is because of the huge WSD based context.
Our semantic approach based on deep enrichment further improves the performance over the individual BoE and BoS. We demonstrated the benefit of judicious combination of semantic features and lexical features drived from both DBpedia and WordNet because they could overcome the ambiguity and the diversity of the vocabulary. The results for the two classification metrics show the utility of expansion features using synsets and concepts together for representing text tweets to improve the quality of categorization. We have increased the accuracy of TC with more than 35% compared to the BoW approach. Finally, the performance differences were statistically significant using our semantic approach for all the classification metrics.

6. Conclusion and Future Work

In this paper, we have presented a novel semantic approach for TC, where the categorization relying only on the ontological knowledge and classifier training is not required. Then, we have proposed a deep enrichment strategy based on feature expansion from external KBs (e.g. DBpedia and Wordnet) to classify tweets into specific categories by using Extended WordNet Domains as a classifier. Our strategy to generate more relevant features for classification by exploiting the semantic concept present in the graph structures of the KBs. The main goal of this work is to show that our deep enrichment using synsets from Wordnet and concepts from DBpedia together improves the quality of TC. Empirical results indicate that this enriched representation of text items based on semantic features can substantially improve the TC accuracy when compared to the lexical features representation (e.g. bag of words (BoW) or bag of entities (BoE)) extracted solely from a Tweet content. Our semantic approach based on the combination of lexical and semantic features has achieved accuracy of more than 80%. The experimental results are especially promising taking into consideration that our approach did not rely on classifier training, and that it can be readily applied to any other set of topics (categories) defined as classification contexts. Our future work aims at integrating our approach into a semantic information retrieval model.

References

[1] Ahmed, A., Rahman, S.S.U., 2017. Dbpedia based ontological concepts driven information extraction from unstructured text. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 8, 411–418.

[2] Aldahawi, H., Allen, S., 2015. An approach to tweets categorization by using machine learning classifiers in oil business, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer. pp. 535–546.

[3] Cano, A.E., Varga, A., Rowe, M., Ciravegna, F., He, Y., 2013. Harnessing linked knowledge sources for topic classification in social media, in: Proceedings of the 24th acm conference on hypertext and social media, ACM. pp. 41–50.

[4] Cotelo, J.M., Cruz, F.L., Enríquez, F., Troyano, J., 2016. Tweet categorization by combining content and structural knowledge. Information Fusion 31, 54–64.

[5] Dilrukshi, I., De Zoysa, K., Caldera, A., 2013. Twitter news classification using svm, in: Computer Science & Education (ICCSE), 2013 8th International Conference on, IEEE. pp. 287–291.

[6] Edouard, A., Cabrio, E., Tonelli, S., Le Thanh, N., 2017. Semantic linking for event-based classification of tweets. International Journal of Computational Linguistics and Applications , 12.

[7] Genc, Y., Sakamoto, Y., Nickerson, J.V., 2011. Discovering context: classifying tweets through a semantic transform based on wikipedia, in: International conference on foundations of augmented cognition, Springer. pp. 484–492.

[8] Gonzalez-Agirre, A., Castillo, M., Rigau, G., 2012. A proposal for improving wordnet domains., in: LREC, pp. 3457–3462.

[9] Hauff, C., Houben, G.J., 2011. Deriving knowledge profiles from twitter, in: European Conference on Technology Enhanced Learning, Springer. pp. 139–152.

[10] Jadidinejad, A.H., 2013. Unsupervised information extraction using babelnet and dbpedia. Making Sense of Microposts (# MSM2013) .

[11] Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A., 2011. Twitter trending topic classification, in: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, IEEE. pp. 251–258.

[12] Li, Q., Shah, S., Ghassemi, M., Fang, R., Nourbakhsh, A., Liu, X., 2016. Using paraphrases to improve tweet classification: Comparing wordnet and word embedding approaches, in: Big Data (Big Data), 2016 IEEE International Conference on, IEEE. pp. 4014–4016.

[13] Magnini, B., Cavaglia, G., 2000. Integrating subject field codes into wordnet., in: LREC, pp. 1413–1418.

[14] McInnes, B.T., Stevenson, M., 2014. Determining the difficulty of word sense disambiguation. Journal of biomedical informatics 47, 83–90.

[15] Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C., 2011. Dbpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th international conference on semantic systems, ACM. pp. 1–8.

[16] Montoyo, A., Palomar, M., Rigau, G., Suárez, A., 2011. Combining knowledge-and corpus-based word-sense-disambiguation methods. CoRR .

[17] Parilla-Ferrer, B.E., Fernandez, P., Ballena, J., 2014. Automatic classification of disaster-related tweets, in: Proc. International conference on Innovative Engineering Technologies (ICIET), p. 62.

[18] De la Pena Sarracén, G.L., . Ensembles of methods for tweet topic classification, in: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017). CEUR Workshop Proceedings. CEUR-WS. org, Murcia (Spain)(September 19 2017).

[19] Robinson, T., 2016. Disaster tweet classification using parts-of-speech tags: a domain adaptation approach. Ph.D. thesis. Kansas State University.

[20] Schulz, A., Guckelsberger, C., Janssen, F., 2017. Semantic abstraction for generalization of tweet classification: An evaluation of incident-related tweets. Semantic Web 8, 353–372.

[21] Selvaperumal, P., Suruliandi, A., 2014. A short message classification algorithm for tweet classification, in: Recent Trends in Information Technology (ICRTIT), 2014 International Conference on, IEEE. pp. 1–3.

[22] Setiawan, E.B., Widyantoro, D.H., Surendro, K., 2016. Feature expansion using word embedding for tweet topic classification, in: Telecommunication Systems Services and Applications (TSSA), 2016 10th International Conference on, IEEE. pp. 1–5.

[23] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M., 2010. Short text classification in twitter to improve information filtering, in: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM. pp. 841–842.

[24] Tare, M., Gohokar, I., Sable, J., Paratwar, D., Wajgi, R., 2014. Multi-class tweet categorization using map reduce paradigm. International Journal of Computer Trends and Technology (IJCTT) 9, 78.

[25] Varga, A., Cano, A.E., Ciravegna, F., et al., 2012. Exploring the similarity between social knowledge sources and twitter for cross-domain topic classification. Proceedings of the Knowledge Extraction and Consolidation from Social Media, CEUR .

[26] Zhao, D., Rosson, M.B., 2009. How and why people twitter: the role that micro-blogging plays in informal communication at work, in: Proceedings of the ACM 2009 international conference on Supporting group work, ACM. pp. 243–252.

[27] Zubiaga, A., Spina, D., Martinez, R., Fresno, V., 2015. Real-time classification of twitter trends. Journal of the Association for Information Science and Technology 66, 462–473.