Features Extraction To Improve Comparable Tweet Corpora Building

Malek Hajjem,Chiraz Latiri

LIPAH resaerch Laboratory, Faculty of Sciences of Tunis
Tunis EL Manar Univeristy,
Campus Universitaire Farhat Hached
B.P. n° 94, 1068 Tunis, Tunisia

Abstract

This paper deals with comparable corpus building from Twitter. We focus on the thematic relevance evaluation process of tweets. In fact, as Twitter microblog is very popular, tweets could be considered as a new data source of comparable corpora. Hence, a possible way to build comparable corpora from Twitter is to extract tweets in two selected languages and sharing a specific topic, in order to construct a multilingual corpus. However, the problem of mining relevant tweets deals with a real challenge: how to only extract the most relevant tweets according to a specific topic from the huge number of collected tweets? In this respect, we propose in this paper an unsupervised machine learning based approach to improve the quality of the collected textual data, in order to identify which messages, *i.e*, tweets, address the specific topic. Several tweets representations are carried out to filter the extracted messages. The main goal of such relevance estimation process is improving the comparability degree between bilingual extracted tweet corpora.

Résumé

Cet article présente une étude expérimentale visant à pallier le problème de la dérive thématique, rencontré lors de la construction d'un corpus comparable à partir de Twitter. En effet, l'exploitation de ce type de microblogs pour la fouille de messages courts est fort intéressante étant donnée la grande masse de données sociales qu'ils offrent. Toutefois, l'extraction de tweets se fait souvent par rapport à des mots clés liés à un thème donné et les résultats d'extraction s'avèrent bruités vu l'existence de tweets qui ne font pas référence au thème du corpus traité. Ainsi, pour éviter cette dérive thématique qui affecte forcément la mesure de comparabilité du corpus bilingue construit, il est indispensable de filtrer les tweets dits non pertinents. À ce titre, nous proposons dans cet article un protocole d'évaluation de la pertinence des tweets collectés en se basant sur une méthode d'apprentissage non supervisée et moyennant différentes représentations de tweets, dans l'objectif d'améliorer la comparabilité du corpus construit.

Mots-clés: Tweet Clustering; Ambiguity Estimation; Twitter mining; Comparable corpora construction; comparability.

1. Introduction and motivations

Twitter is a micro-blogging service on the Web where users create status messages (called *tweets*) to express their thoughts and reactions in multiple languages. This micro-blogging is currently one of the most popular sites of the Web with at least 304 millions monthly active users [1]. In one hand, a generous microblogs comments become publicly visible making text mining in social networks a challenge ground for a variety of research efforts but, on the other hand, the writing style of microblogs messages tends to be quite informal. So, different linguistic phenomena in the textual data are observed. In this respect, several problems for different NLP applications are mentioned, especially available NLP tools which have become inappropriate to deal with such orthographic variations. In this paper, we focus on the task related to relevance evaluation process of tweets. While considering the affluence of textual data in web social resources, we defend that using text mining techniques in social networks can serve as an interesting way for multilingual corpora building, especially comparable corpora construction (13; 4). However, the extracted data appears usually raw and noisy. Moreover, we notice that tweets extraction process is only based on keywords related to a specific topic or subject and we observe that the crawling result includes many irrelevant thematic tweets. We devote this article to contribute to the Twitter's text preprocessing issue.

In our research work, thematic tweet filtering represents an important step of comparable corpora building process from Twitter. We investigate a cluster-based approach for classifying tweets in order to avoid this **thematic drift**, by mean of proposing different tweet corpus representation for clustering purposes. As a first proposal, a text based only approach mainly founded on the vocabulary corpus is handled. Then, we propose to combine with the original vocabulary corpus an external resource to overcome the short length of the tweet messages which hampers the mining process.

The remainder of the paper is organized as follows: a brief literature review is given in section 2. Then, section 3, introduces the problem statement and presents a critical discussion. In section 4, the first strategy to filter our comparable tweet corpus is introduced. Section 5 is dedicated to the second tweet clustering strategy based on an external resource. Experimental results are described in section 6. The conclusion and future work are finally presented in section 7.

2. Related work

2.1. Tweet Corpora construction

Multilingual corpora are either parallel corpora that contain source text and their translations, or comparable corpora which are collections of documents in the same or in different languages made up of texts dealing with the same subject. As parallel corpora are more expensive to obtain, building comparable corpora was the best alternative. Moreover, the development of internet makes web the principal source of multilingual corpora, especially comparables ones (10). Recently, the emergence of microblogs attracted much attention from researchers. Due to the restrictions on their distribution (14),textual analysis which conduct to the construction

[1] http://www.blogdumoderateur.com/chiffres-twitter/

of corpora based on social networks platforms and microblogs still restricted. Nevertheless, some studies addressed the corpora construction issue based on social networks platforms and microblogs such as (8) where authors proposed an approach to build a monolingual detection event corpus. In (15), authors have dealt with automatic mining of parallel corpora from Sina microblogs, the most popular social media in china. In (5), mining multilingual tweets has mainly relied on Cross-Lingual-Information-Retrieval techniques. Some others studies such as (13; 4; Fraisse and Paroubek), focused on the construction a comparable corpus from Twitter. Recent research (7) discusses the important challenges of collecting the data collection from social media.

2.2. Tweets Classification

Tweets classification issue has been addressed in different contexts. The literature offered several supervised and unsupervised approaches for this classification task. In (17), authors presented a basic classification technique for twitter messages, based on a Naive Bayes classifier, to detect whether a tweet is related or not to a given company. Some other proposals defend that using an unsupervised technique for short text grouping is more efficient. In fact, short text classification requires an important number of training example to achieve significant accuracy. In this area, authors in (18) proposed a general unsupervised framework to explore events from tweets. The filtering step conducts a lexicon-based approach to separate tweets that are event-related from those that are not. In unsupervised case one possible approach is modifying the term weighting technique. The idea of weighting is to give more weight to the terms of higher importance. In this context, authors in (16) have proposed an alternative technique to $tf \times idf$ term weighting in short documents based on normalized cut (Ncut) method (12). Another possible approach is to assume that short text don't provide enough shared context. Research in (9) indicates that even highly related Twitter messages often have very little overlapping on the word level and need to be extended. In (11; 2), authors mentioned the need of incorporating some external knowledge such as introducing external corpus as an additional knowledge which enables the use of external semantics. Indeed, Ferragina et al. (2) proposed a method based on entity disambiguation for short texts using Wikipedia. In the same way, Seifzadeh et al. (11) worked on a novel representation of short-text segments, enriched with information about correlation between terms.

3. Problem statement

In this section, we will describe the textual data collection extracted from Twitter and we will discuss the thematic quality of the resulting comparable corpus.

3.1. Building comparable corpus from Twitter

Twitter is widely used as an easy, fast and convenient new broadcasting tool. Thanks to its official APIs, we can get access to the tweet data servers. This prevents us from using unsuitable existing technologies of web crawling to extract data from Twitter. In our case two kind of APIs were used: search api[2], a part of Twitter's Rest API which allows queries against popular

[2] https://dev.twitter.com/rest/public/search

French keywords	Translation	Arabic keywords	Transliteration
Printemps arabe	Arab spring	الرّبيع العربيّ	Alrrabyς Alςrbyy
Révolution	Revolution	الثّورة	Alθwrħ
Syrie	Syria	سوريَا	swryA
Egypte	Egypt	مصر	mSr
Révolution tunisienne	Tunisian revolution	الثّورة التونسيّة	Alθwrħ Altwnsyyħ

Table 1: Sample of keywords Arabic/French
Note: The transliteration consists on writing Arabic with latin characters to help non Arabic speakers to read Arabic.
اَ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
A b t θ j H x d r z s š S D T Ď ς γ f q k l m n h w y
and the additional letters: ’ ء, Â أَ, Ǎ إِ, Ā آ, ŵ ؤُ , ŷ ءى , ħ ة, ý ى.

Tweets. The second one was Topsy Otter API [3], mainly used to broaden the corpus. Unlike Search API, this one makes possible to retrieve historical tweets that had the most visibility according to a given topic. We collected tweets about Arab Spring Revolutions in French and Arabic language. This process is a basic information retrival approach. It requires querying for messages containing relevant keywords to the assigned topic. Different keywords including multiple ways of spelling are used. (*Cf.* Table 1). Redundancy was eliminated by deleting retweets, as they do not generally add any new information. A language based on filtering was also performed using a language detection library[4] to remove non-French tweets.(*Cf.* Table 2)

	French	Arabic	Average word number	Publishing period
Search Api corpus	20025	20023	5	May 2013 to September 2013
Topsy Api corpus	150000	202752	6	December 2010 to February 2011

Table 2: Characteristics of the constructed comparable corpora (Post Pre-processing)

3.2. Critical discussion

Comparable corpora construction from Twitter is based basically on using ad-hoc keywords pre-defined by researchers. The main challenge is how much the resulting keywords searches find tweets relevant to the assigned topic ? Indeed, the tweets corpora resulting from these ad-hoc searches are constrained to the keywords selected by **human intuition** (6). The limitation of this approach is that it ingnores the context of these words. Thus, nothing guarantee that all microblog messages regarding to certain topic are retrieved or that all retrieved messages are related to the Arabe Spring (5). As a result, tweets corpus does not necessarily represents a suitable tweets collection, covering relevant thematic tweets according to the searched topic.

[3] http://topsy.com/

[4] https://code.google.com/p/language-detection/

In fact, the keywords ambiguity and the use of an uncorrelated large list of words give rise that some irrelevant tweets would be retrieved. Even if these tweets contain one of pre-defined keywords may not fit with the specific topic such as depicted in Table 3. This problem makes building comparable corpora from twitter as a challenging task, since thematic homogeneity is one of comparability criteria. In this paper, we try to address this question through a tweets classification task. The key feature of our proposal is mining through tweets that contain a set of keywords, whether or not they are related to the given assigned topic. Hence, messages related to Arab Springer Revolution will be separated from others irrelevant tweets. The main goal of such relevance estimation process is improving the comparability degree between bilingual extracted tweet corpora.

T1	The army will not allow ”an against-*revolution*” in Egypt	T
T2	Freebox *revolution* : download is free today	F
T3	Peaceful *Revolution* in Iceland #revolution #anticapitalism	F
T4	New MacBook Pro ranges & Mabook. It is a *revolution*... or not!	F

Table 3: Translated Tweets containing the keywords “revolution”

4. Towards improvement of features selection

Tweets classification for thematic filtering purposes can be handled as a binary classification task. Tweets that are relevant to the given topic represent the positive category and those that are irrelevant constitute the negative category. Two ways are offered to group similar tweets, namely : supervised classification or unsupervised classification. To decide, which one is more appropriate to deal with a large tweets corpus, we need to study the size, the nature and the distribution of the collected tweets data. In our work, both relevant and irrelevant tweets are randomly distributed, which makes the use of supervised classification unsuitable. It is worth noting that building a non noisy training data set from twitter for a supervised classifier is a hard task which needs a large team of human experts such as done in the IR evaluation campaigns. Thus, unsupervised classification seems the more adequate way to the filtering non relevant tweets. Our idea is to use a non-hierarchical clustering technique, namely K-MEANS algorithm, to group relevant tweets according to the Spring Arabic Revolution topic in a same cluster. Several possibilities of vector’s tweets representation for clustering purposes are proposed. We aim to achieve the best representation, able to improve the comparability on the bilingual built corpus.

4.1. Preprocessing steps

Different steps of preprocessing are handled before the clustering process. First, we have eliminated stopwords such as auxiliary verbs and articles. Then, special characters (*i.e.*, names of users, punctuation, smileys, etc) and numbers, were excluded from each collection in order to keep a pure textual data. Composed words are also extracted. We notice that the stemming phase is avoided because the tweet vocabulary is already limited. Finally, before proceeding with tweet representation, we have created a training and test datasets to be used in next steps.

In our work, the training dataset represents the major part of the corpus. It involves: (1) the n best words extraction related to Arab Spring Revolution topic from the tweets vocabulary in terms of frequency; (2) the weights of these words. Regarding the test dataset, it is a hand-labeled dataset mainly used to evaluate the accuracy of the tweet clustering process. The annotation phase of this dataset aims to binary annotate the tweet corpus. In fact tweets have to be separated into 2 classes: First class represents the tweets related to the topic assigned (positive class) and second class represents tweets that are not related to the topic assigned (negative class).

In all we have used a half positive labeled set and a half negative labeled set. The two sets were extracted from the constructed corpus from twitter based on a human decision process. We notice that all the clustering process was performed in this article for the French language tweets.

In the following, we address the feature selection issue after preprocessing.

4.2. Filtering based on TF-Mesure

Tweets are represented as a normalized TF (Term Frequency) features vectors, where a tweet represents a data point in n-dimensional space with n is the size of the considered corpus tweet vocabulary, extracted from the training corpus. A $tweet_i$ is represented as a vector $(v_1, v_2, \ldots, v_n)$ where v_j is the TF for of j^{th} word x_j in $tweet_i$, as follows:

$$\begin{pmatrix} v_1 = TF(x_1) \\ v_2 = TF(x_2) \\ v_3 = TF(x_3) \\ \ldots \\ v_n = TF(x_n) \end{pmatrix}$$

where $TF(x_n)$ is a normalized term frequency in the represented tweet.

4.3. Filtering based on weighted representation

The second representation relies on the same principle of the previous presentation, except that each frequency is weighted by its weight in the whole training corpus as follows :

$$\begin{pmatrix} \alpha(x_1) \cdot TF(x_1) \\ \alpha(x_2) \cdot TF(x_2) \\ \alpha(x_3) \cdot TF(x_3) \\ \ldots \\ \alpha(x_n) \cdot TF(x_n) \end{pmatrix}$$

where $\alpha(x_j)$ is the normalized frequency of the word x_j in the training corpus.

The goal of this representation is to give more importance to the terms of higher weight compared to others words. In case of short text, term frequency of most of the words is limited in our documents (mostly 1, rarely 2 or 3), the weighted representation vector would actually boil down to a pure normalized frequency of words in the training dataset which is a reduced but a significative value.

4.4. *Filtering based on inverted rank*

Third representation relies on the same principle of the two previous representations, except that each frequency is weighted by its inverted rank in the whole training corpus as follows:

$$\begin{pmatrix} \beta(x_1) \cdot TF(x_1) \\ \beta(x_2) \cdot TF(x_2) \\ \beta(x_3) \cdot TF(x_3) \\ \cdots \\ \beta(x_n) \cdot TF(x_n) \end{pmatrix}$$

where $\beta(x_j) = \frac{1}{r(x_j)}$ and $r(x_j)$ represents the rank of x_j in the training corpus.

5. Tweet clustering using external web corpus

The most challenging issue related to short clustering text is the sparsity of a such text representation. In fact, Twitter limits the length of each Tweet to 140 characters. Therefore, traditional techniques for calculating text similarity give usually scores close to zero. We notice that short documents even similar ones have very few terms in common. In order to overcome the short length of tweet messages, we suggest to enrich the tweets representation with additional knowledge related to correlation between vocabulary terms. In other words, if a term x_j is absent in a given $tweet_i$, but its co-occurrent term appears in this tweet, we can conclude that this tweet is semantically related to the assigned topic. In this case, the co-occurrent term could replace the absent original term in the representation. In order to achieve this goal, we propose to enhance tweet clustering process using co-occurrence words extracted from an external thematic web corpora.

5.1. *Thematic web corpus construction and preprocessing*

Thematic web corpus construction is a crucial phase in this tweet clustering representation. This task aims to capture a set of URLs believed to be relevant to a specific theme or topic. In fact, the web corpus must be in the same thematic area (*cf.*, Figure 1) of the tweets corpus since it will be used to extract correlations between tweet corpus vocabulary and the web corpus vocabulary. To be sure that this corpus contains only relevant documents, this task is conducted on two steps. First, we performed a *google* search to generate web link relevant to Arab Spring Revolution topic through the open source jsoup java html parser[5]. We note that search options take into account the date proximity of web generated pages, even if this constraint will restrict the web corpus size. Actually, our aim is to ensure the maximun homogeneity between the two corpora content. Then a boilerplate html remover was performed to extract just the text search result. As a final result, we obtain 1000 web documents about Arab spring, mostly related to web newspaper articles such as *Liberation*, *Le*

[5] https://jsoup.org/apidocs/

Figaro, *Le monde*, etc. More informations about the web constructed corpus are presented in Table 4. Finally, as in all text processing application, the first step was morphological analysis. This step is carried out using the morpho-syntactic analyzer TreeTagger[6]. Tokenization and stop words elimination were the sub-tasks that were followed in our method. Only nouns and verbs where considered to extract co-occurrences which will be described in the next section.

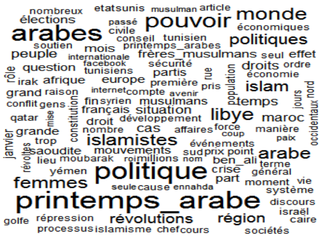

Figure 1: Cloud for the 100 most frequently words in the web corpus

Web corpus	
Articles	1000
Sentences	14229
Average sentences/article	19
Average words/article	116
Number of words	25562
Vocabulary	11730

Table 4: French Web corpus statistics

5.2. Words co-occurrence extraction from web corpora

To extract co-occurrences, we suggest a method based on Mutual Information (MI) measure. This measure allows to compute the association degree between two given words and then to make up lists of the most correlated words (1). Regarding MI, we need to represent the web corpus in a Vector Space Model (VSM). We obtain a term document matrix (TDM) where there are w columns (unique considered words) and t documents; each cell measures the importance of the word within each document. Values are set to the frequencies. In TDM, terms that have a large sparsity values are not considered. Then, a co-occurrences matrix is generated to compute the associations between words as follows :

$$MI(x,y) = p(x,y) \times \frac{p(x,y)}{p(x) \times p(y)} \tag{1}$$

5.3. Filtering based on co-occurrences representation

This representation relies on the same principle of the third one, except that absent terms in a given tweet will be represented by their best co-occurrence words. Frequency of the best co-occurrence word is weighted by the same inverted rank of the original word in the training corpus as follows :

$$\begin{pmatrix} \beta(x_1) \cdot TF(x_1) \\ \beta(x_2) \cdot TF(x_2) \\ \beta(x_3) \cdot TF(x_3) \\ \dots \\ \beta(x_n) \cdot TF(x_n) \end{pmatrix}$$

with

$$TF(x_j) = \begin{cases} TF(x_j) - \alpha \ \ if \ \ TF(x_j) > 0 \\ \frac{\alpha}{N} \times TF(\text{Best-co-occurrence}(x_j)) \\ \ \ if \ \ TF(x_j) = 0 \end{cases}$$

where N is the tweet size

α is a smoothing parameter

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

6. Experiments and results

Conducted experiments on the French tweets corpus aim to compare the accuracy of these different representations. The goal is to achieve the best representation able to extract the best partition. Once obtained, this list of the *n* best words will be our first statistical features for the clustering process.

We used the WEKA machine learning toolkit[7]. Classes to clusters validation was conduct in this evaluation process. In this mode, The machine learning toolkit this method basically does the classification trough clustering in two steps: 1) first it ignores the class attribute which are ideally produced by human judges and generates the clustering. 2) Then during the test phase it assigns classes to the clusters, based on the majority value of the class attribute within each cluster.

Thus, to evaluate the performance of the clustering process, we use the standard classification performance metric used in information retrieval field and text classification studies, namely : precision (*P*), recall (*R*) and the F-measure (*F-measure*) calculated based on the confusion matrix and computed as follows:

$$P = \frac{Tp}{Tp + Fp} \qquad R = \frac{Tp}{Tp + Fn} \qquad F\text{-}mesure = \frac{2 \times P \times R}{P + R}$$

where Tp is the number of tweets correctly assigned to this class; Fp is the number of tweets incorrectly assigned to this class and Fn is the number of tweets incorrectly rejected to this class.

We compare the performances of different representations using K-MEANS classifier. Experiments are applied on 20% of our dataset representing the test corpus. For all the representations, we will apply K-MEANS classifier while incrementing our statistics features (*i.e.*, incrementally from 2 to 150 most frequent words). Table 5 summarizes the clustering performance for the three proposed representations.

	Precision	**Recall**	**F-measure**
TF-Measure Rep	0.616	0.523	0.565
Weighted Rep	0.649	0.588	0.617
Inverted-rank Rep	0.513	0.966	0.670
Co-occurrence Rep	0.632	0.836	0.719

Table 5: Classifier's Accuracy

The results show that with the 6 best words, we achieve the best F-measure for the three first representations only based on modifying the term weighting technique. Table 5 and Figure 2 highlight that the representation based on the inverted rank weighting maximizes the F-measure in this case. Regarding the co-occurrence representation, Table 5 shows that this representation

[7] http://www.cs.waikato.ac.nz/ml/weka/

achieved the best clustering accuracy based on F-measure score. An F-measure equal to 0.719 is obtained by using the additional best co-occurrence words of the 150 best first words in the vector representation. This means that our features expanded from external web corpus were able to avoid some spareness terms. It is worth noting that since the tweets are short and belong to a single topic, expanding them with only their associated words, can not efficiently overcome the spareness problem. In fact, many words have the same best associations words and extracted co-occurrences from web corpus were missed.

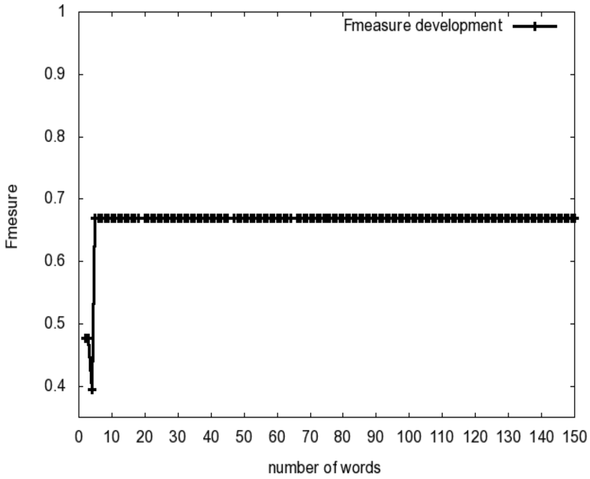

Figure 2: Evolution of F-measure with respecting the number of considered features (words)

7. Conclusion and Ongoing Work

In this paper, we have investigated the utility of an experimental study on the thematic ambiguity estimation of a tweets corpus. The originality of this contribution doesn't lie in the classification process but specifically on the final objective which is improving the similarity degree in comparable tweets corpus. This contribution focused on a statistical features selection to achieve a efficient filtering of the collected tweets. As a work in progress, we are trying to improve the clustering process by adding other features especially those relying on using an extracted thematic space through LDA approach from a large external corpus.

References

Coccaro, N. and Jurafsky, D. (1998). Towards better integration of semantic predictors in statistical language modeling. In *Proceedings of ICSLP-98*, pages 2403–2406.

Ferragina, P. and Scaiella, U. (2012). Fast and accurate annotation of short texts with wikipedia pages. *IEEE Softw.*, 29(1):70–75.

Fraisse, A. and Paroubek, P. Twitter as a comparable corpus to build multilingual affective lexicons. In *proceedings of the 7th International Workshop on Building and Using Comparable Corpora at LREC 2014 (BUCC 2014), Reykjavik, Iceland*, pages 17–21.

Hajjem, M., Latiri, C. C., and Slimani, Y. (2014). Twitter as a multilingual source of comparable corpora. In *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia, Kaohsiung, Taiwan, December 8-10, 2014*, pages 342–345.

Jehl, L., Hieber, F., and Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 410–421, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph, K., Landwehr, P. M., and Carley, K. M. (2014). An approach to selecting keywords to track on twitter during a disaster. In *Proceedings of the 11th International ISCRAM Conference*, ISCRAM 2014.

Mayr, P. and Weller, K. (2016). Think before you collect: Setting up a data collection approach for social media studies. *arXiv preprint arXiv:1601.06296*.

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information; Knowledge Management*, CIKM '13, pages 409–418, New York, NY, USA. ACM.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). 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*, WWW '08, pages 91–100, New York, NY, USA. ACM.

Saad, M., Langlois, D., and Smaïli, K. (2014). Cross-lingual semantic similarity measure for comparable articles. In *Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings*, pages 105–115.

Seifzadeh, S., Farahat, A. K., Kamel, M. S., and Karray, F. (2015). Short-text clustering using statistical semantics. In Gangemi, A., Leonardi, S., and Panconesi, A., editors, *WWW (Companion Volume)*, pages 805–810. ACM.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905.

Trabelsi, M., Hajjem, M., and Latiri, C. (2014). Building comparable corpora from social networks. In *The 7^{th} Workshop on Building and Using Comparable Corpora*.

Weller, K. and Kinder-Kurlanda, K. (2015). Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research?

Xing, H., Yang, M., Qi, H., Li, S., and Zhao, T. (2013). Mining parallel corpus from sina microblog. In *Proceedings of the 2013 International Conference on Asian Language Processing*, IALP '13, pages 99–102, Washington, DC, USA. IEEE Computer Society.

Yan, X., Guo, J., Liu, S., Cheng, X.-q., and Wang, Y. (2012). Clustering short text using ncut-weighted non-negative matrix factorization. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2259–2262, New York, NY, USA. ACM.

Yerva, S. R., Miklós, Z., and Aberer, K. (2011). What have fruits to do with technology? : The case of orange, blackberry and apple. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 48–58, New York, NY, USA. ACM.

Zhou, D., Chen, L., and He, Y. (2015). An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2468–2475.