Disambiguating company names in microblog text using clustering for online reputation management

Desambiguando nombres de compañías en microblogs usando agrupamiento para el manejo de su reputación online

Fernando Pérez-Tellez

SOCIAL MEDIA RESEARCH GROUP
INSTITUTE OF TECHNOLOGY TALLAGHT DUBLIN
IRELAND
fernandopt@gmail.com

John Cardiff

SOCIAL MEDIA RESEARCH GROUP
INSTITUTE OF TECHNOLOGY TALLAGHT DUBLIN
IRELAND
john.cardiff@itt.dublin.ie

Paolo Rosso

NATURAL LANGUAGE ENGINEERING
LAB - PRHLT RESEARCH CENTER
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
SPAIN
proso@dsic.upv.es

David Pinto

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
MEXICO
dpinto@cs.buap.mx

Recibido: 25-VI-2013 / **Aceptado:** 24-III-2014

Abstract

Twitter is used by millions of users to publish brief messages (tweets) with the purpose of sharing experiences and/or opinions about a product or service. There is a clear need for systems that can mine these messages in order to derive information about the collective thinking of twitterers (e.g. for opinion or sentiment analysis). Tweet analysis is a very important task because comments, opinions, suggestions, complaints etc. can be used for marketing strategies or for determining information on a company's reputation. For this purpose, it is necessary to automatically establish whether a tweet refers to a company or not, when the company name is ambiguous. This task is not a straightforward keyword search process as there may be multiple contexts in which a name can be used. The aim of this study is to present and compare four different approaches which improve the representation of short texts for better performance of the clustering task that determine whether a given tweet refers to a particular company

or not. For this purpose, we have used a variety of enriching methodologies based on term expansion via the semantic similarity hidden behind the lexical structure, in order to improve the representation of tweets and as a consequence the performance of the task. We have used two different tweet datasets of company names which contain different levels of ambiguity. The results are promising although they highlight the difficulty of this task.

Key Words: Clustering of tweets, opinion analysis, disambiguation, online reputation management.

Resumen

Twitter es utilizado por millones de personas con la finalidad de publicar mensajes cortos con el propósito de compartir experiencias y/u opiniones acerca de un determinado producto o servicio. Existe una clara necesidad de crear sistemas que sean capaces de analizar estos mensajes a fin de derivar información sobre el pensamiento colectivo de las personas que los publican. El análisis de los *tweets* se ha convertido en una tarea muy importante para las grandes compañías, debido a que los comentarios, sugerencias y quejas pueden ser usados como estrategias de mercadotecnia o para determinar la reputación de cierta compañía. Entre otras tareas, es necesario construir métodos que permitan determinar, de forma automática, cuando un *tweet* se refiere a una compañía o no, en el caso de que el nombre de la compañía sea ambiguo. El objetivo de este trabajo es presentar y comparar cuatro diferentes aproximaciones, para la desambiguación de nombres de compañías mediante métodos de agrupamiento. Para este propósito hemos propuesto una variedad de metodologías de enriquecimiento basada en la expansión de términos vía la similitud semántica escondida detrás de la estructura léxica, todo esto con el objetivo de mejorar la representación de los *tweets* y como consecuencia el desempeño de la tarea de categorización. En los experimentos se han usado dos conjuntos de *tweets*, los cuales contienen nombres de compañías con diferentes niveles de ambigüedad. Los resultados obtenidos son prometedores y al mismo tiempo demuestran la dificultad de la tarea.

Palabras Clave: Agrupamiento de *tweets*, análisis de opinión, desambiguación, manejo de reputación *online*.

INTRODUCTION

Twitter¹ - the microblog platform that allows users to publish brief messages of less than 140 characters- is a Web 2.0 application which offers a new mode of user interaction. It has become an important channel through which users can share their experiences or opinions about a product, service or company, and companies are taking advantage of this medium as part of their marketing strategies. It has been estimated by Complete² that the use of Twitter has been drastically increased from 2009 to 2012, reaching up to 45 million unique visitors; however the increase in 2012 was not as significant as in previous years. In 2012 and 2013 Twitter has been

the third most popular online social networking platform according to Statista Inc.³. These facts demonstrate the high popularity of this new publishing medium and the evident importance that it provides to the end reputation management strategies of companies.

Internet users commonly look for recommendations on platforms such as Twitter on products or services before buying them. The recommendations or feedback from consumers about a product are based on personal experiences. It is critical that companies be aware of such online conversations, so that they can react quickly to any negative discussion and use it in order to provide better products and services. From this phenomenon a new necessity appears, i.e., the automatic analysis of the company's reputation. Online reputation management - the monitoring of media and the detection and analysis of opinions about an entity is becoming an active area of research, and such systems have become a necessity for small companies, mid-size businesses, large corporations and organisations alike (Kempe, Kleinberg & Tardos, 2003).

This research work presents a first stage of the online reputation management process by identifying messages that refer to a particular company on the Twitter platform. We demonstrate that a term expansion methodology can improve the representation of tweets from a clustering perspective. We present a comparison of similar approaches in order to determine which one produces the best improvement from the clustering perspective.

In this work, we present an approach with some variants in order to categorise tweets which contain a company name, into two clusters corresponding to those which refer to a company and those which do not. Providing a solution to this problem will allow companies to be aware of immediate user reaction to their products or services, and thereby manage their reputations more effectively (Milstein, Chowdhury, Hochmuth, Loric & Magoulas, 2008).

The rest of this paper is organised as follows. Section 1 describes the related work and the problem description. Section 2 presents the data set used in the experiments. Section 3 explains the approaches and techniques used in this research work. Section 4 shows the experiments, the obtained results and a discussion of them. Finally, the last Section presents the conclusions.

1. Problem description

Nowadays consumers are using the Web technologies such as Twitter to elicit opinions about a product or service before make buying decisions. For a company to monitor its reputation and apply the proper actions to keep the consumers happy is very important because bad reputation can result in big losses. Based on this we are interested in discriminating between messages that refer to a company, and those that do not. This is particularly pertinent in cases where the company name also has

a separate meaning in the English language, for instance, the term ‘delta’ can refer to the fourth letter of the Greek alphabet, or alternatively to a tract of sediment deposited at the mouth of a river. However, the ‘delta’ term is also used by the well-known American airline with the same name which is the company that we are interested in finding the related microblogs entries. We can find many such cases of ambiguity in company names (e.g. ‘palm’, ‘ford’, ‘borders’) amongst others. In our work, we regard a company name as ambiguous if the term that comprises its name can be used in different contexts. An example can be seen in Table 1 where the word ‘borders’ is used in the context of a company (row 2) and as the boundary of a country (row 3). We are particularly interested in the problem of clustering company entries extracted from Twitter. But considering that a company name may be very ambiguous and the number of characters in Twitter is restricted to 140, it makes the clustering of company names a very complicated task. In order to provide a better picture of the problem under discussion, we can consider the examples of the tweets in Table 1. As we may see, all the tweets contain the word ‘borders’ but not all of them are related to the ‘Borders’ book store company. The tags provided in the second column of the table for each tweet (‘Yes’ or ‘No’), indicate whether or not the tweet is associated with the company.

Table 1. Samples of tweets which contain the ‘borders’ word.

Tweet	Related to the company
DONT TELL ME EVEN BORDERS ALSO NVR SELL THE FIX!!!!!!	Yes
excessively tracking the book i ordered from borders.com. kfgjdfkgyfd.	Yes
With a severe shortage of manpower, existing threat to our borders, does it make any sense to send troops to Afghanistan? @centerofright	No
Help Haiti!Purchase a full size Skin so soft product and 50cents will be donated to redcross and doctors without borders for Haiti relief	No
33% O_ Borders Coupon : http://wp.me/pKHuj-qj	Yes

The size of a microblog text is an intrinsic characteristic and also a drawback for clustering approaches. Classical term weighting scheme such as TF-IDF (Jones, 1972) will usually fail, since the term frequencies will be very low. Moreover the small vocabulary size in conjunction with the writing style makes the task more difficult. Tweets are written in an informal style, and may also contain misspellings or be grammatically incorrect.

In order to improve the representation of the tweets we have proposed an approach based on an expansion procedure (enriching semantic similarity hidden behind the lexical structure). In this research work, we demonstrate that a term expansion methodology, as presented in this document, can improve the representation of the tweets from a clustering perspective, and as a consequence obtain better defined groups in the clustering task.

2. Related work

2.1. Tweet classification

The categorisation of tweets is a topic that has become of great interest for computational researchers due to the impact that this type of data analysis could have in studies of company marketing. However, research works dealing with the problem of word ambiguity (in this case, to determine whether a word refers to a company or not) have only been studied in literature recently (Amigo, Artiles, Gonzalo, Spina & Liu, 2010).

We describe briefly here the work which is related to the problem of clustering short texts related to companies, and in particular, those works in the field of categorisation of tweets and clustering of short texts. In Sankaranarayanan, Samet, Teitler, Lieberman and Sperling (2009) an approach is presented for binary classification of tweets (class ‘breaking news’ or other). The class ‘breaking news’ is then clustered in order to find the most similar news tweets, and finally a location of the news for each cluster is provided. Tweets are considered short texts as mentioned in Sriram, Fuhry, Demir and Ferhatosmanoglu (2010) where a proposal for classifying tweets is presented. This work addressed the problem by using a small set of domain-specific features extracted from the author’s profile and the tweet text itself. They claim to effectively classify the tweet to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages. Therefore, it is important to analyse some techniques for categorisation of short texts. Authors in (Yamashita, Sato, Oyama & Kurihara, 2013) study the commonality between friends in Twitter platform; the assumption being that users are following each other because they have many different commonalities so they can be clustered on the basis of the nature of the commonality. Users are defined as nodes and the ‘following’ relationship among users as edges.

2.2. Named Entity Disambiguation

Named Entity Disambiguation is the process of identifying and classifying phrases/terms in a text that may refer to people, places or organisations. The task of disambiguating named entities has been study for long texts but researchers have just started studying methods to address classification of short text in the last years. For instance, (Ferragina & Scaiella, 2010) present a system capable of annotating short text such as snippets of search-engine results and tweet. The system uses the linking structure of Wikipedia pages and the anchor text of the links to deduct possible senses for a given short fragment of text. The system resolves the ambiguity and polysemy in the potentially many available anchor text (linked pages) by finding the collective agreement among them via scoring functions. In Meij, Weerkamp and de Rijke (2012) a method is proposed to add semantics to tweets in order to facilitate social media

mining by identifying concepts that are semantically related and generating links to Wikipedia pages. This approach involves two steps: the generation of a ranked list of candidate concepts and the selection of the proper candidates. Machine learning techniques are used to refine the candidate Wikipedia concepts.

2.3. Online reputation management

We present related approaches that deal with the management of collective intelligence of crowds and communities published on the web, in particular to the online reputation management. Recent work (Fan, Ju & Xiao, 2013) focuses its discussion on how reputation affects revenue, prices, transaction volume, survival likelihood and how seller manage their reputation in a large Chinese e-commerce website. The authors have found that seller reputation has positive impact on established sellers which is not the case with new sellers. In other words, new sellers sacrifice short-run benefits of reputation in pursuits for the long return to reputation and established sellers get better return to reputation.

The works which have attempted a solution on the tweet categorization task as part of Task 2 of the WePS-3 evaluation campaign⁴ for the online reputation management are summarised in (Amigo et al., 2010). The authors describe five approaches to tackle this task. The best system reported is the LSIR-EPFL approach (Yerva, Zoltán & Aberer, 2010) which uses additional resources for classification such as Wordnet⁵, meta-data from the webpage, Google results and user feedback. The ITC-UT system (Yoshida, Matsushima, Ono, Sato & Nakagawa, 2010) was the next system best which uses an initial supervised categorisation step composed by a set of rules based on Part-of-Speech tags and Named Entity recognition process to predict the ambiguity of the company name. The SINAI system (García-Cumbreras, García-Vega, Martínez-Santiago & Peréa-Ortega, 2010), uses external resources or tools such as Named Entity Recognisers, Wikipedia⁶ and DBpedia⁷ information and then manual generated rules are applied. In contrast to these approaches, we propose the alternative of not using any training data or additional resources due to the fact that we consider it difficult to find linguistic resources for all kind of domains. Our proposals use information included in the dataset which exploits the intrinsic relationship among terms. In addition, we also present an alternative which may use external limited information to enrich these relations. The reason for using this external resource is only for comparison purposes to show that the improvement after our main proposal is not adding additional value to our approach. The approaches presented in WePS use classification approaches due to the fact that it is well-known that supervised approaches will produce better results than unsupervised ones. In this case, a training subset was provided. Even though clustering produces weaker results than classification approaches, we have chosen to use an unsupervised approach based on the assumption that training sets do not exist for all domains. We want to demonstrate that clustering results can be improved by adapting and augmenting the clustering algorithm with other techniques.

In the experiments carried out in our research work, we considered datasets related to user-generated contents, such as company tweets, in contrast with the systems described above with uses rating score approaches from the consumers. We have found that the initial problem of online reputation management is the process of identifying the relevant information about a particular company. In the next section, we describe the dataset used in our experiments.

3. Dataset description

We have generated our datasets from one task of a well-recognised international competition named the WePS-3 evaluation campaign⁸. Two tasks concerning the problem of Web entity search were proposed in the WePS-3 evaluation campaign. The first task was related to Web People Search, the problem of person name ambiguity; whereas the second task was related to Online Reputation Management for organisations, i.e., it was focused on the problem of organisation (company) name ambiguity. The corpora were obtained from the trial and training data sets of Task 2 of this evaluation campaign. The trial corpus contains entries for 17 (English) and 6 (Spanish) organisations; whereas the training data set contains 52 (English) organisations. It was labelled by five annotators, each of whom voted for the most appropriated label. The ‘true’ label means that the tweet is associated to a company, whereas the ‘false’ one means that the tweet is not related to any company, and the ‘unknown’ label is used to indicate that the annotators were unable to make a decision.

In this paper we are conducting an experiment on clustering company tweets and, therefore, a more homogeneous corpus than the one provided by the WePS-3 competition is desirable. For this purpose, we have modified the corpus to select those company tweets with information written in English, and considering only the true and false tweets, i.e., we do not consider the unknown label. Furthermore, the subset used in the experiments includes only those 20 companies (see Table 2) with a sufficient number of positive and negative samples (true/false), i.e., at least 20% of the total items must be in each category. The reason for doing this is because we are proposing an expansion methodology which needs minimal accessible information to be available to perform the enrichment process, in other words, it is difficult to use our approaches with limited information (close to zero) in each of the tweet categories (true/false) for the expansion process.

Table 2. Statistics of tweets used in the experiments.

Company	Total**	True**	False**	Vocabulary Size	Number of Words	Average Words*	Minimum Words*	Maximum Words*
Bestbuy	98	24	74	704	1441	14.70	6	22
Borders	94	25	69	665	764	12.29	2	20
Delta	96	39	57	584	1178	12.27	5	20
Ford	97	62	35	700	1241	12.79	2	22
Leapfrog	96	70	26	393	1262	13.14	3	20
Opera	98	25	73	671	1208	12.32	1	25
Overstock	94	70	24	613	1301	13.84	3	22
Palm	99	28	71	762	1406	14.20	4	22
Southwest	99	39	60	665	1348	13.61	4	21
Sprint	94	56	38	624	1138	12.10	3	22
Armani	415	312	103	2325	6357	13.64	2	23
Barclays	419	286	133	2217	6715	14.10	2	24
Bayer	371	228	143	2105	6136	13.63	3	22
Blockbuster	437	306	131	2309	5595	11.75	3	21
Cadillac	427	271	156	2449	5880	12.19	2	24
Harpers	437	142	295	2356	6042	12.20	2	23
Lennar	99	74	25	438	1324	13.37	5	21
Mandalay	435	322	113	2085	6012	12.42	2	22
Mgm	431	177	254	1977	6545	13.63	2	24
Warner	99	23	76	596	1302	13.15	4	20

** Number of tweets

* Number of words in tweets

Finally, each selected company subset must contain at least 90 labelled tweets as minimum number of tweets associated, as we normally can find tweets about companies meeting this criterion. In Table 2, we present a detailed description of the corpus features. In the following section we present and compare four different approaches for dealing with this problem. Our rationale for having these different approaches is to show the variations in terms of improvement of the results using each methodology. We also want to show that the previous steps to the clustering process can help in the improvement of the representation of the data and produce better results.

4. Clustering the tweet Dataset

The objective of this research work is to cluster tweets which contain an ambiguous word that can refer to a company name and divide them into two groups, those that refer to the company and those that refer to a different topic. We approach this problem by introducing and subsequently evaluating four different methodologies that use term expansion. The term expansion approach has the goal of improving the representation of the data from a clustering perspective, in other words, we may obtain better clustering results following this process.

The term expansion of a set of documents is a process for enriching the semantic similarity hidden behind the lexical structure. Although the idea of term expansion has been previously studied in literature (Banerjee & Pedersen, 2002) (Pinto, Rosso & Jimenez-Salazar, 2010) we are not aware of works in which it is applied to microblog texts.

In order to establish the difficulty of clustering company tweets, we split the 20 companies group into two groups that we hypothetically considered easier and harder to cluster. The first group is composed of 10 companies with generic names, i.e., names that are expected to be very ambiguous (words that appear in a dictionary with different meanings). The second group contains specific names which are considered to be less ambiguous (words that can be used in a limited number of contexts or words that do not appear in a dictionary). We expect the latter group will be easier to categorise than the former. In Table 3, we see the distribution of the two groups.

We have selected the K-means clustering method for the experiments carried out in this work. The reason is that it is a well-known method, it produces acceptable results and our approaches may be compared with future implementations. This clustering method (MacQueen, 1967) is one of the most popular iterative clustering algorithms, in which the number of clusters ‘k’ has to be fixed a-priori. K-means chooses ‘k’ different centroids and, thereafter, it associates each item to the nearest centroid. ‘K’ new centroids are then re-calculated and the process is repeated iteratively. For the purposes of this case study, we have established the parameter ‘k’ to be equal to two, those pertaining to companies and those not related. In order to construct the similarity matrix which will be used by K-means for constructing the clusters, we used a tweet representation based on ‘tf-idf’ (see Eq 1) with the similarity between tweets calculated by means of the cosine measure (see Eq 2).

Table 3. Types of Company names.

<i>Generic Company Names</i>				
Bestbuy	Borders	Delta	Ford	Southwest
Leapfrog	Opera	Overstock	Palm	Sprint
<i>Specific Company Names</i>				
Armani	Barclays	Bayer	Blockbuster	Lennar
Cadillac	Harpers	Mandalay	Mgm	Warner

The Term Frequency and Inverse Document Frequency (tf-idf) is a statistical measure of weight often used in natural language processing to determine how important a term is in a given corpus, by using a vectorial representation. The importance of each term increases proportionally to the number of times this term appears in the document (the frequency), but is offset by the frequency of the term in the corpus. We will refer to the ‘tf-idf’ as the complete similarity process of using the ‘tf-idf’ weight and a special similarity measure proposed by Salton in (Salton, Wong

& Yang, 1975) for the Vector Space Model, which is based on the use of the cosine between two vectors representing the documents.

The 'tf' component of the formula is calculated by the normalised frequency of the term, whereas the 'idf' is obtained by dividing the number of documents in the corpus by the number of documents which contain the term; and then taking the logarithm of that quotient. Given a corpus D and a document d_j ($d_j \in D$), the 'tf-idf' value for a term t_i in d_j is obtained by the product between the normalised frequency of the term t_i in the document d_j (tf_{ij}) and the inverse document frequency of the term in the corpus ($idf(t_i)$) as follows:

$$tf - idf = tf_{ij} * idf(t_i) \quad \text{Eq1}$$

Where

$$tf_{ij} = \frac{tf(t_i, d_j)}{\sum_{k=1}^{|d_j|} tf(t_k, d_j)} \quad \text{and} \quad [idf(t_i)] = \log \left(\frac{|D|}{|d: t_i \in d, d \in D|} \right)$$

Document Frequency (DF) is a technique that assigns the value $DF(t)$ to each term 't', where $DF(t)$ means the number of texts in a collection, where 't' occurs. The assumption in this technique is that low frequency terms will rarely appear in other documents, therefore these terms will not contribute on the classification of a text, in other words, the technique is based on the fact that rare terms are not discriminative important for determining the target cluster of a document.

Each document can be represented by a vector where each entry corresponds to the 'tf-idf' value obtained by each vocabulary term of the given document. Thus, given two documents in vectorial representation, $\vec{d}_i$ and $\vec{d}_j$, it is possible to calculate the cosine of the angle between these two vectors as follows:

$$\cos_0 = (\vec{d}_i, \vec{d}_j) = \frac{\vec{d}_i \cdot \vec{d}_j}{\|\vec{d}_i\| \|\vec{d}_j\|} \quad \text{Eq2}$$

The similarity matrix is then constructed on the basis of the above formulae, i.e., for each possible pair of tweets we need to calculate how similar they are by using the cosine measure. Once the similarity matrix is calculated, we may proceed with the clustering step.

4.1. Self-Term Expansion Methodology

The Self-Term Expansion Methodology (S-TEM) (Pinto et al., 2010) comprises a twofold process: the Self-Term Enriching Technique, which is a process of replacing terms with a set of co-related terms, and a Term Selection Technique with the role of identifying the relevant features. In the particular case of the S-TEM methodology, we use only the information being clustered to perform the term expansion, i.e., no external resource is employed. In Figure 1, we illustrate the main steps of this methodology. In general terms, the first step takes the information (tweets) in order to generate the co-occurrence list, and based on it, we estimate the relevant relations in order to generate the expanded corpus. After this step, the selection process obtains the most discriminative information for each category and the new expanded corpus is sent to the clustering process.

Figure 1. Self-Term Expansion Methodology.

4.1.1. Self-Term Enriching Technique

The Self-Term Enriching Technique improves the representation of short documents by using a term enriching (or term expansion) procedure. No external resources are employed because we consider that it is quite difficult to identify appropriate linguistic resources for every particular kind of text. Moreover, we intend to exploit intrinsic properties of the same corpus to be clustered in an unsupervised way. In other words, we take the same information that will be clustered to make the term expansion; that is the reason of this technique to be called self-term expansion.

The technique consists of replacing terms of a document with a set of co-related terms. A co-occurrence list is calculated from the target data set by applying Pointwise Mutual Information (PMI).

PMI provides a degree of relationship between two words; however, the level of this relationship must be empirically adjusted for each task based on the length of documents and dataset. In this work, we established empirically a PMI value equal or greater than two to be the best threshold. In other experiments (Pinto, 2008) dealing with longer and more formal texts, a threshold of six was used; however in documents correlated terms are rarely found because of the low term frequencies.

The Self-Term Enriching Technique is defined formally in (Pinto et al., 2010) as follows: Let $D = \{d_1, d_2, \dots, d_n\}$ be a document collection with vocabulary $V(D)$. Let us consider a subset of $V(D) \times (D)$ of correlated terms as $RT = \left\{ \begin{matrix} t_i, t_j \\ t_j, t_i \end{matrix} \mid t_i, t_j \in V(D) \right\}$. The RT expansion of D is $D' = \{d_1, d_2, \dots, d_n\}$, such that for all $d_i \in D$, it satisfies two properties:

- 1) if $t_i \in d_i$ then $t_j \in d'_i$, and
- 2) if $t_i \in d_i$ then $t'_j \in d_i$ with $[(t_i, t'_j)] \in RT$. If RT is calculated by using the same target data set, then we say that D' is the Self-Term Enriched version of D.

4.1.2. Term Expansion Technique

The function of this technique is to identify the best features and reduce the noise produced by the enriching technique. However, it is also useful to reduce the computing time of the clustering algorithms. In particular, we have used Document Frequency (Jones, 1972). The Document Frequency technique assumes that low frequency terms will rarely appear in other documents; therefore, they will not have significance in the prediction of the class of a document.

4.2. Term Expansion Methodology – Wiki (TERM-Wiki)

This technique also called TEM-Wiki; it is based on the two previously presented techniques (Term Enriching Technique - Wiki and the Term Selection Technique). The main application of this methodology is in clustering microblog texts. In particular, it is determining if a microblog which contains a term (name of a company such as 'apple') refers to the well-known company with the same name or to something else, for instance the fruit.

The main variation of this technique is that before the creation of the list of co-occurrence from the same corpus, we have added valuable information to the corpus which is extracted from Wikipedia related to the company before the estimation of important relationships between words and the generation of the co-occurrence list. The hypothesis is that the information of the organisation or company will highlight the relevant relations between the terms in the collection and increase the frequency of the related terms which may be included in the co-occurrence list. In this case, the value of PMI was kept equals to two for the estimation of the relevant relationships among terms.

Each company subset was enriched with the corresponding company information provided by Wikipedia. We have extracted the information of all the companies contained in our corpus and we have enriched the particular company information with its corresponding Wikipedia information. The aim of this approach is to improve the representation of microblogs by using self information contained in the corpus and additional information extracted from the external resource (in this case, Wikipedia). Figure 2 illustrates the main components of this methodology. Term Enriching Technique (indicated by broken lines on the left) incorporates information from Wikipedia.

The Term Enriching Technique – Wiki (TEM-Wiki) is defined formally as follows: Let $D = \{d_1, d_2, \dots, d_n\}$ be a document collection with vocabulary $V(D)$, $WK = \{wk_1, wk_2, \dots, wk_n\}$ be a Wikipedia document and the concatenation of D and WK as $DWK = D \bullet WK$ which consist of all the strings of the form $d \bullet wk$ with vocabulary $V(DWK)$. Let us consider a subset of $V(D) \times V(DWK)$ of correlated terms as $RTW = \left\{ \begin{smallmatrix} t_i, t_j \\ t_i, t_j \end{smallmatrix} \in V(DWK) \right\}$. The RTW expansion of D using the information contained in D plus Wikipedia information is $D' = \{d_1, d_2, \dots, d_n\}$, such that for all $d_i \in DWK$, it satisfies the two properties 1) and 2) from Section 4.1.1. If RTW is calculated by using the same target data set, then we say that D' is the Term Enriched version of D . After the Term Enriching Technique using the same corpus plus adding information taken from Wikipedia the Term Expansion Technique is applied.

Figure 2. Term Expansion Methodology – Wiki.

4.3. Full Term Expansion Methodology (TEM-Full)

In this methodology, we have used two techniques: the Full Enriching Technique and the Term Expansion Technique. For the former, we expand only the ambiguous word (the company name) with the terms that co-occur within the same company subset, in other words, if a co-occurrence is detected with the company name, the term is candidate to be included even the value of it is not high. For this, we have used minimal restrictions such as the level of frequency and the value of the PMI of the terms. The pairs in the co-occurrence list must produce a value greater than zero of PMI and the terms must have at least a frequency of two in the whole corpus, all of these in order to be considered for the co-occurrence list. It is important to mention that we have used the Term Selection Technique in order to select the most discriminative terms for the categories. The process is shown in Figure 3. Note that this expansion process does not use an external resource at all. We believe that due to the low term frequency and the shortness of the data (tweets), it is better to include most of the terms that co-occurs in the subset of a company and provide more information to the enriching process. We expect to provide better well-defined groups to help in the clustering process.

Figure 3. Full Term Expansion Methodology.

The formal definition of Full Term Enriching Technique (TERM-Full) is as follows: Let $D = \{d_1, d_2, \dots, d_n\}$ be a document collection with vocabulary $V(D)$ and w a particular term in D where $|\{D\}| \leq |\{W\}|$. Let us consider a subset of $V(D) \times V(D)$ of correlated terms as $RT = \left\{ \begin{matrix} t_i, w \\ t_i, w \end{matrix} \in V(D) \right\}$. The RT expansion of D is $D' = \{d'_1, d'_2, \dots, d'_n\}$, such that for all $d_i \in D$, it satisfies the two properties 1) and 2) from Section 4.1.1. If RT is calculated by using the same target data set, then we say that D' is the

Full Term Enriched version of D. After the Full Enriching Technique is applied using the same corpus with no external resources the Term Expansion Technique is applied.

4.4. Full Term Expansion Methodology with a Formaliser (TEM-Full+F)

This methodology consists of two parts as the previous methodologies: the Full Enriching Technique and the Term Expansion Technique. A key feature of this methodology is the formalisation of microblog documents, as users tend to write their comments using abbreviations, due to the length restriction of 140 characters. We have used an abbreviation dictionary⁹ that contains 5,173 abbreviations commonly used in microblogs, tweets and short messages. We have replaced the abbreviations with their formal equivalent throughout the corpus before the expansion process. We have called this step of replacing terms in microblogs documents as the formalisation process. After the formalisation step, the expansion is performed but it is only applied to the ambiguous word (the company name) and words which highly co-occur with it. These latter words were selected as they appear in frequencies with the ambiguous word in positive tweets (related to the companies). We consider that this kind of word may help making the correct decision during the clustering process because they are highly related to the company name/ambiguous word.

Table 4. Example of words expanded using this methodology.

Armani	men, exchange, accessories, gorgio, purchase, shipping, emporio
Borders	online, off, book, web, bookstores, product, group
Delta	airlines, crew, jet, flight, airplane
Mandalay	bay, resort, casino, hotel, vegas, beach, suite, accommodations, pool
Opera	software, technology, developers, interface, web, browser
...	

In this particular methodology, we have expanded the ambiguous term and the related terms that highly co-occur with the company name. In Table 4, we present some samples of these terms which were enriched. The terms were empirically extracted as they are limited number and we have seen that limited number of terms produce acceptable results due to the fact that the terms are highly coo-related to the ambiguous company names.

The ambiguous (company names) and related words are enriched in the corpus only, the enriching process is done by using the full enriching technique, in other words, we enriched on the basis of values of PMI greater than zero and that the frequency of the terms higher than two. All the terms filtered were used to concatenate the ambiguous word and the related words thus we expand the corpus. The final step is the selection of the discriminative information on the basis of the Term Selection Technique. The process of this methodology is shown in Figure 4. On contention

that if we formalise the text as a pre-processing step and we also expand the terms that co-occur with the company name, it will provide better information for the clustering process.

The formal definition of Full Term Enriching Technique with a Formaliser (TERM-Full+F) is as follows. Let $D'=\{d'_1, d'_2, \dots, d'_n\}$ be a document collection after a formalisation process, with vocabulary $V(D)$ and W a proper subset of D with vocabulary $V(W)$. Let us consider a subset of $V(D) \times V(W)$ of correlated terms as $RT=\left\{ \begin{matrix} t_i, w_j \\ t_i, w_i \end{matrix} \in V(D) \right\}$. The RT expansion of D is $D'=\{d'_1, d'_2, \dots, d'_n\}$, such that for all $d_i \in D$, it satisfies the two properties 1) and 2) from Section 4.1.1. If RT is calculated by using the same target data set, then we say that D' is the Full Term Enriched version of D . After the Full Enriching Technique with the variation of being applied to the company names and the list of related terms the Term Expansion Technique is applied.

Figure 4. Full Term Expansion Methodology with a Formaliser.

5. Experimental Results

The aim of these experiments is to verify whether or not an enriching procedure would help in improving the task of clustering company tweets. Therefore, we have tested the four different methodologies proposed in the previous section over the datasets presented. In order to compare the performance of the different approaches, we have calculated a baseline in which we use clustering with K-means, without any enriching procedure. The obtained results using the different methodologies proposed are compared in Table 5, the bold text represents the cases when the result is better than the baseline.

There is a clear improvement in most of the approaches by comparison with the baseline. This indicates that the enriching procedure is a good technique of document representation. The results have shown improvement in most of the cases. However, we got two of the methodologies with disimproved the baseline (S-TEM and TEM-Full) using the ‘warner’ company subset. It happened because this company subset contains the largest difference between positive and negative posts (23-76 respectively), fact that is difficult to find in real scenarios because most of the recognised companies in the Twitter platform are very likely to have more than only 30 related posts. In our case, the enriching procedure needs posts containing terms that co-occur highly so they can help in finding relations. TEM-Full+F generated the same value as the baseline and the only methodology which improved the baseline was TEM-Wiki. It makes sense due to the fact that the positive information to enrich the corpus for this company was limited and the Wikipedia information helped to find better coo-relations.

Table 5. A comparison of each methodology with respect to one baseline using the F-measure (Bold text represents the cases when the result outperformed the baseline).

Companies	Methodologies				
	S-TEM	TEM-Wiki	TEM-Full	TEM-Full+F	Baseline
Generic Company Names Subset					
Bestbuy	0.68	0.69	0.74	0.75	0.62
Borders	0.64	0.68	0.73	0.72	0.60
Delta	0.76	0.78	0.71	0.70	0.61
Ford	0.60	0.60	0.67	0.65	0.64
Leapfrog	0.69	0.69	0.71	0.63	0.63
Opera	0.70	0.66	0.73	0.74	0.70
Overstock	0.68	0.68	0.66	0.72	0.58
Palm	0.65	0.65	0.72	0.70	0.62
Southwest	0.62	0.64	0.67	0.72	0.64
Sprint	0.60	0.60	0.67	0.65	0.64
Specific Company Names Subset					
Armani	0.66	0.66	0.73	0.70	0.62
Barclays	0.65	0.65	0.72	0.72	0.55
Bayer	0.65	0.72	0.71	0.70	0.63
Blockbuster	0.66	0.71	0.71	0.71	0.66
Cadillac	0.61	0.61	0.69	0.69	0.61
Harpers	0.65	0.65	0.68	0.68	0.63
Mandalay	0.68	0.68	0.74	0.84	0.64
Mgm	0.72	0.72	0.54	0.75	0.69
Lennar	0.97	0.97	0.72	0.97	0.96
Warner	0.61	0.74	0.54	0.67	0.67

‘Ford’ and ‘Sprit’ subsets showed similar behaviour in all the methodologies because the low term frequency and the relations detected among terms were poor so that they were difficult to be used in S-TEM and TEM-Wiki, on the other hand, the full enriching and the formalisation of TEM-Full and TEM-Full+F process made the difference because they included specific coo-related terms so they produced better input for the clustering process.

We would also like to mention that the good results presented in companies such as ‘Lennar’ or ‘Southwest’ were obtained because the vocabulary of these microblog documents showed small vocabulary size and also low overlapping vocabulary between the two categories (positive and negative) and, therefore, the clustering process could find well-delimited groups.

Figure 5. Comparison among the methodologies proposed (average).

The TEM-Full methodology has shown the best performance with the corpus of generic company names. In this case, we have expanded only the ambiguous word (the name of the company), whereas the TEM-Wiki methodology performed well with the corpus of specific company names. TEM-Wiki uses information extracted from Wikipedia for enriching the meaning of the company names. The TEM-Full+F methodology has shown good performance in both corpora. It takes advantage from the formalisation step and the enriched words that co-occur with the ambiguous word in the positive examples. We would like to highlight TEM-Full+F methodology because it did not disimprove the baseline in any case even if the information for one category (positive or negative) was limited for that we consider that this methodology good option. We have observed that, regardless of whether or not we are using an external resource, we may improve the representation of company tweets for the clustering task. Figure 5 presents the performance of the four approaches by using average values obtained from both subsets. It is provided for better understanding of the results. We consider it significant that even though the TEM-Full+F methodology does not obtain the maximum performance, it provides constant improvement in the majority of cases.

In Figure 6 and Figure 7, we show the performance of the two approaches (TEM-Full and TEM-Wiki) which were obtained with the ‘generic’ and ‘specific’ company name corpus, respectively. In particular, the TEM-Full methodology has shown a good improvement in the performance of clustering generic company names (see Figure 6). On the other hand, in Figure 7 it is possible to see that the TEM-Wiki methodology outperformed the baseline for the most of the company names. The best result has been achieved when the text was formalised and the words that co-occur in positive tweets were enriched (TEM-Full+F methodology) for both kinds of corpora (see Figure 8).

Figure 6. The TEM-Full methodology applied to the ‘generic’ company name corpus.

Figure 7. The TEM-Wiki methodology applied to the ‘specific’ company name corpus.

This methodology has not shown disimproved results in any instance and it produces good results in most cases. Although the term expansion procedure has been shown to be effective for improving the task of clustering company tweets, we believe that there is still room for improving the obtained F-Measure values. It is important to note that the F-measure value obtained by the TEM-Wiki approach is

slightly different in comparison with the values obtained by the TEM-Full and TEM-Full+F. This fact may lead us to conclude that the clustering of company tweets is a very difficult task. Even though our results cannot be compared directly with the one presented in the WePS-3 campaign our average results are between 0.60 and 0.70 of F-measure, which means a 10% improvement over the baseline.

Figure 8. The TEM-Full+F methodology applied to both kinds of corpora.

CONCLUSION

Clustering short text corpora is a difficult task. Since tweets are by definition as short texts, the clustering of tweets is also a particularly complex problem. Furthermore, due to the nature of writing style of these kinds of texts –informal writing style (a poor grammatical structure) with many out of vocabulary words– this kind of data typically causes most clustering methods to obtain poor performance.

The main contribution of this paper has been to propose and compare a number of innovative approaches for improving clustering in microblog texts; we have tested them on a dataset containing company name references, in order to evaluate the effectiveness of the methodology for the purpose online reputation management. We have found that the idea of term expansion helped in most cases, particularly for the best approach (TEM-Full+F) which never disimproved the baseline. Moreover, although the idea of term expansion has been previously studied in literature, this is the first time of which we are aware that it has been applied to microblog texts. We introduced four methodologies for enriching term representation of tweets. We expected that these different representations would lead classical clustering methods, such as K-means, to obtain a better performance than when clustering the same dataset and the enriching methodology is not applied.

In order to validate the difficulty of clustering company tweets, we constructed two datasets, one with specific and other with generic company names, which we

hypothetically considered easier and harder to be clustered, respectively. By observing the obtained results it is not possible to demonstrate that one dataset is easier to be clustered than the other and indeed both seem quite difficult. In particular, we consider that (TEM-Wiki) performed well on the former dataset while another methodology (TEM-Full) obtained the best results on the latter dataset. However, we have considered the TEM-Full+F methodology suitable for both kinds of corpora, which does not require any external resource. S-TEM and TEM-Full+F are completely unsupervised approaches. They construct a thesaurus from the same dataset to be clustered and, thereafter, the approaches use this resource for enriching the terms. Moreover, TEM-Full+F methodology takes advantage of the formalisation of the text that users avoid sometimes for the length restriction on the Twitter platform. On the other hand, TEM-Wiki uses information from Wikipedia that is introduced by human beings.

We approached the problem of clustering microblog texts from different angles (we proposed four different methods for the representation of tweets based on term expansion). We believe that this helps to give a more complete idea of how difficult and challenging the WePS task of Online Reputation Management for organizations is. We have used these different approaches in order to provide a better comparison and to have an understanding of the difficulty of this task. On the basis of the results presented, we can say that using this particular data, the unsupervised methodology (TEM-Full+F) has shown slightly better results than the rest of the methodologies presented. The best results were obtained due to formalisation of the text and the inclusion of all the information that co-occurs in the corpus of a particular company. It is important to say that this methodology (TERM-Full+F) showed a good performance regardless of the shortness and low term frequency of the data analysed.

REFERENCES

- Amigo, E., Artiles, J., Gonzalo, J., Spina, D. & Liu, B. (2010). WePS-3 Evaluation Campaign: Overview of the Online Reputation Management Task. In *2nd Web People Search Evaluation Workshop (WePS 2010)*. Padova, Italy: CLEF 2010 Conference.
- Banerjee, S. & Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. *Proceedings of the CICLing 2002 Conference* (pp. 136-145). LNCS: Springer-Verlag.
- Fan, Y., Ju, J. & Xiao, M. (2013). *Losing to Win: Reputation Management of Online Sellers*. Univeristy of Arizona and University of Oklahoma [on line]. Retrieved from: http://www-personal.umich.edu/~yingfan/reputation_FJX.pdf
- Ferragina, P. & Scaiella, U. (2010). TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). *Proceedings of the 19th ACM international conference on Information and knowledge management*. (pp. 1625-1628). Toronto, ON, Canada: ACM.
- García-Cumbreras, M., García-Vega, M., Martínez-Santiago, F. & Peréa-Ortega, J. (2010). SINAI at WEPS-3: Online Reputation Management. In *2nd Web People Search Evaluation Workshop (WePS 2010)*. CLEF 2010 Conference.
- Jones, S. K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.
- Kempe, D., Kleinberg, J. & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *KDD'03: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 137-146). USA: ACM.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). Berkeley: University of California Press.
- Meij, E., Weerkamp, W. & de Rijke, M. (2012). Adding Semantics to Microblog Posts. *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 563-572). Seattle, Washington, USA: ACM.
- Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B. & Magoulas, R. (2008). *Twitter and the micro-messaging revolution: Communication, connections, and immediacy-140 characters at a time*. O'Really Radar Report.
- Pinto, D. (2008). *On Clustering and Evaluation of Narrow Domain Short-Text Corpora*. Unpublished doctoral dissertation, Universidad Politécnica de Valencia, Valencia, Spain.

- Pinto, D., Rosso, P. & Jimenez-Salazar, H. (2010). A self-enriching methodology for clustering narrow domain short texts. *The Computer Journal*. Oxford University Press.
- Salton, G., Wong, A. & Yang, C. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613-620.
- Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M. & Sperling, J. (2009). Twitterstand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 42-51). USA: ACM.
- Sriram, B., Fuhry, D., Demir, E. & Ferhatosmanoglu, H. (2010). Short text classification in twitter to improve information filtering. *The 33rd ACM SIGIR'10 Conference* (pp. 42-51). USA: ACM.
- Yamashita, T., Sato, H., Oyama, S. & Kurihara, M. (2013). Classification of Twitter Users Based on Following Relations. *Proceedings of the International MultiConference of Engineers and Computer Scientists, I*. Hong Kong.
- Yerva, S. R., Zoltán, M. & Aberer, K. (2010). It Was Easy, when Apples and Blackberries Were only Fruits. In *2nd Web People Search Evaluation Workshop (WePS 2010)*. Padua, Italy: CLEF 2010 Conference.
- Yoshida, M., Matsushima, S., Ono, S., Sato, I. & Nakagawa, H. (2010). ITC-UT: Tweet Categorization by Query Categorization for On-line Reputation Management. In *2nd Web People Search Evaluation Workshop (WePS 2010)*. CLEF 2010 Conference.

NOTES

- 1 <http://twitter.com>
- 2 <http://siteanalytics.compete.com/twitter.com/>
- 3 <http://www.statista.com/statistics/71336/top-10-social-media-websites-in-the-us-by-market-share/> - Most popular social media websites in the United States in January 2013, based on share of visits.
- 4 <http://nlp.uned.es/weps/weps-3/guidelines/40-guidelines-for-the-weps-3-on-line-reputation-management-task>
- 5 <http://wordnet.princeton.edu/>
- 6 <http://www.wikipedia.org/>
- 7 <http://dbpedia.org/>
- 8 WePS3: searching information about entities in the Web, <http://nlp.uned.es/weps/>, February 2010
- 9 <http://noslang.com/dictionary>