On the Difficulty of Clustering Company Tweets

Fernando Perez-Tellez
Institute of Technology Tallaght Dublin
Tallaght, Dublin 24
Dublin, Ireland
fernandopt@gmail.com

David Pinto
Faculty of Computer Science
Benemérita Universidad Autonóma de Puebla
Puebla, Mexico
dpinto@cs.buap.mx

John Cardiff
Institute of Technology Tallaght Dublin
Tallaght, Dublin 24
Dublin, Ireland
John.Cardiff@ittdublin.ie

Paolo Rosso
Natural Language Engineering Lab, ELIRF, DSIC,
Universidad Politécnica de Valencia
Valencia, Spain
prosso@dsic.upv.es

ABSTRACT

Twitter is a new successful technology of the Web 2.0 genre which is used by millions of people and companies to publish brief messages ("tweets") with the purpose of sharing experiences and/or opinions about a product or service. Due to the huge amount of information available in this type of technology, there is a clear need for new 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 can be used as marketing strategies or for determining information on a company's reputation. For this purpose, it is necessary to establish whether a tweet refers to a company or not, which is not a straightforward keyword search process as there may be multiple contexts in which a name can be used. The aim of this work is to present and compare a number of different approaches based on clustering that determine whether a given tweet refers to a particular company or not. For this purpose, we have used an enriching methodology in order to improve the representation of tweets and as a consequence the performance of the clustering company tweets task. The obtained results are promising and highlight the difficulty of this task.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation

SMUC'10, October 30, 2010, Toronto, Ontario, Canada.

Keywords

Clustering of tweets, Opinion analysis

1. INTRODUCTION

In recent years, the Internet has emerged as an important tool of communication and socialization. As a part of this evolution of Internet, new Web 2.0 applications such as wikis, weblogs and social networks have appeared. The Twitter[1] -microblog that allows users to publish brief message updates- is one of these applications and has become an important channel in which users can share their experiences or opinions about a product, service or a company. In general, companies have taken advantage of this medium for developing marketing strategies. It has been estimated by Complete[2] that the use of Twitter has been drastically increased in the last five years. Actually, in [9] it has been reported that in 2008 Twitter has grown at a rate of 1382% in terms of the number of tweets sent. This fact shows us the high popularity of this new publishing medium and the evident potential that it could provide to the marketing of companies.

In general terms, Twitter is a service that allow anyone to send anything to anybody in 140 characters or less. Twitter has gained huge popularity since it was launched [3]. It has been attracting increasing interest from the research community. It is clear that this kind of instantaneous publishing medium could be useful for the companies that can access automatically this kind of information regarding themselves. This requires the ability to filter tweets in order to know which are related to the companies and which are not. We consider clustering to be a good approach to this task because it can be unsupervised, and there is no need of training data set that in many cases may not be available.

In this paper we present a first approach to the categorization of tweets which contain a company name, into two clusters corresponding to those which refer to the company

[1]http://twitter.com
[2]Complete.com - Site profile for twitter.com. http://siteanalytics.compete.com/twitter.com/, March 2010

Table 1: Examples of "True" and "False" tweets that contains the *Borders* word

TRUE	DONT TELL ME EVEN BORDERS ALSO NVR SELL THE FIX!!!!!!!
TRUE	excessively tracking the book i ordered from borders.com. kfjgjdfkgjfd.
FALSE	With a severe shortage of manpower, existing threat to our borders, does it make any sense to send troops to Afghanistan? @centerofright
FALSE	Help Haiti!Purchase a full size Skin so soft product and 50cents will be donated to redcross and doctors without borders for Haiti relief
TRUE	33% Off Borders Coupon : http://wp.me/pKHuj-qj

and those which do not. Providing a solution to this problem will allow companies to access to the immediate user reaction to their products or services, and thereby manage their reputations more effectively [10].

Online reputation management systems has become a necessity for small companies, mid-size business, large corporations and organizations alike [6]. These systems may take advantage of the relatively easy manner of posting tweets of the Internet users. In this way, this research work presents a first stage for online reputation management system by identifying the tweets that are relevant for a particular company on the Twitter platform. We demonstrate that a term expansion methodology can improve the representation of tweets from a clustering perspective, then we present a comparison of similar approaches in order to determine which one produces the best improvement from the clustering perspective for our particular problem. We are interested, for instance, in finding tweets that include the word "Apple", but only those tweets that refer to the very-well known computer company.

The rest of this paper is organized as follows. Section 2 describes the related work and the problem description. Section 3 presents the data set used in the experiments. Section 4 explains the approaches and techniques used in this research work. Section 5 shows the experiments, the obtained results and a discussion of them. Finally, Section 6 presents the conclusions.

2. PROBLEM DESCRIPTION

Twitter has become a critical source of information and reputation management. We are interested in discriminating entries that correspond to a company from those that do not refer to a company, in particular where the company name also has a separate meaning in the English language (e.g. *delta*, *palm*, $ford$, *borders*). In this research work, we refer as a ambiguous company name to a word or words that compound a company name that can be used in different contexts as it is shown in Table 1.

The size of the tweet is an intrinsic characteristic and also a drawback for the clustering approaches. Classical term weighting scheme such as TF-IDF [5] will usually fail, since the term frequencies will be very low (frequency 1 or 2). Moreover the small vocabulary size in conjunction with a informal 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.

In this research work we state that a term expansion methodology, presented in this paper, can improve representation of the tweets, and as a consequence the performance of the clustering task. In addition, we believe that specific company names -names that can not be found in a dictionary- such as *Lennar* or $Warner$ may be easier to be identified than generic company names such as *Borders*, *Palm* or *Delta*, because of the high ambiguity of the latter company names.

The categorization 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) rarely have been studied in literature. We describe briefly here the work which is in some way related to the problem of clustering short texts related to companies. In particular those works in the field of categorization of tweets and clustering of short texts.

In [14] 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 [15] where a proposal for classifying tweets is presented. This work addressed this 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 categorization of short texts. In [2], for instance, a method is proposed for improving the clustering of short texts (news or blog feeds) by enriching their representation using information from Wikipedia. A more detailed discussion on the literature of categorization of short texts may be found in [11].

In this paper we are 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 in discussion, some examples of tweets are shown in Table 1. They are part of the original subset that corresponds to *Borders* (an English bookstore) As we may see all the tweets contain the word *borders* but not all of them are related to the book store company. The tags provided at the beginning of each tweet ($true$ or $false$) indicate whether or not the tweet is associated with the company.

In the experiments carried out in this paper we considered datasets related with user-generated contents, such as company tweets. The description of the corpora used is given in the following section.

3. DATASET DESCRIPTION

We focused these experiments in one task of a well recognized international competition named WePS-3 evaluation campaign[3]. 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, i.e., it was focused on the problem of person name ambiguity; whereas, the second task was related to Online Reputation Management for organizations, i.e., it was focused on the problem of organization (company) names ambiguity. In particular, the corpora were obtained from the *trial* and *training* data sets of the task 2 of this evaluation campaign. The *trial* corpus of task 2 contains entries for 17 (English) and 6 (Spanish) organisations; whereas the *training* data set contains 52 (English) organisations. The WePS-3 corpus of task 2 was labeled by five annotators, they have 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.

We must consider that in this paper we are approaching a preliminary 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 selected those company tweets with information written in English considering only the *true* and $false$ tweets, i.e., in the experiments carried out we do not considered the *unknown* label. Furthermore, the subset used in the experiments includes only those 20 companies 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. Finally, each selected company must contain at least 90 labeled tweets, which was the minimum number of tweets associated to a company found in the collection.

In Table 2 we present a detailed description of the corpus features such as the number of *true* and $false$ tweets, the average length of the tweets (average number of words),the minimum and maximum number of words contained in tweets, and the vocabulary size of each company tweet set.

In the following section we present different approaches for dealing with this problem.

Table 2: Statistics of company tweets used in the experiments.

Company	T/F	◇	□	△	◯	▽
Bestbuy	24/74	704	1441	14.70	6	22
Borders	25/69	665	764	12.29	2	20
Delta	39/57	584	1178	12.27	5	20
Ford	62/35	700	1241	12.79	2	22
Leapfrog	70/26	393	1262	13.14	3	20
Opera	25/73	671	1208	12.32	1	25
Overstock	70/24	613	1301	13.84	3	22
Palm	28/71	762	1406	14.20	4	22
Southwest	39/60	665	1348	13.61	4	21
Sprint	56/38	624	1138	12.10	3	22
Armani	312/103	2325	6357	13.64	2	23
Barclays	286/133	2217	6715	14.10	2	24
Bayer	228/143	2105	6136	13.63	3	22
Blockbuster	306/131	2309	5595	11.75	3	21
Cadillac	271/156	2449	5880	12.19	2	24
Harpers	142/295	2356	6042	12.20	2	23
Lennar	74/25	438	1324	13.37	5	21
Mandalay	322/113	2085	6012	12.42	2	22
Mgm	177/254	1977	6545	13.63	2	24
Warner	23/76	596	1302	13.15	4	20

T/F - No. of true/false Tweets,
◇ - Vocabulary size,
□ - No. of words,
△ - Average words in Tweets,
◯ - Minimum number of words in Tweets,
▽ - Maximum number of words in Tweets.

4. CLUSTERING COMPANY TWEETS

The purpose of this research work is to bring together (cluster) tweets that contain a possible company entity into two groups, those that refer to the company and those that refer to a different topic. We approach this problem by introducing and, thereafter, evaluating four different methodologies that use term expansion. The term expansion of a set of documents is a process for enriching the semantic similarity hidden behind the lexical structure. Even if the idea of term expansion has been previously studied in literature [12, 4, 1, 11], in this paper, we evaluate the performance of four different approaches for term enriching in the task of clustering company tweets, which are presented as follows:

1. Self-Term Expansion Methodology (S-TEM): A self-term enriching technique and a term selection technique are used.

2. Term Expansion Methodology - Wiki (TEM-Wiki): It consists of a enriching technique that uses the same corpus plus additional information extracted from Wikipedia. A term selection technique is also applied.

3. Term Expansion Methodology with Positive examples - Wiki (TEM-Positive-Wiki): The TEM-Wiki methodology is used for enriching only "positive" tweets (those that really refer to companies). The terms were also enriched (expanded) with information extracted from Wikipedia. Even if this methodology is not completely unsupervised because in a real case we do not know what tweet is "positive", we were interested in knowing the its performance for comparison purposes.

4. Full Term Expansion Methodology (TEM-Full): In this case, we expand the ambiguous word (the word that is possible referring to a company) with all those words that co-occur with it in the same class of the corpus. A term selection technique is also applied.

In order to validate the difficulty of clustering company tweets, we split out the 20 companies group into two groups that we hypothetically considered easier and harder to be clustered. 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). On the other hand, the second group contains specific names which are considered to be less ambiguous (words that can be used in limited number of contexts or words that do not appear in a dictionary). We expect the latter group will be easier to

[3]WePS3: searching information about entities in the Web, http://nlp.uned.es/weps/, February 2010

Table 3: Types of Company names

Generic Company Names			
BestBuy	Borders	Delta	Ford
Leapfrog	Opera	Overstock	Palm
Southwest	Sprint		
Specific Company Names			
Armani	Barclays	Bayer	Blockbuster
Cadillac	Harpers	Mandalay	Mgm
Lennar	Warner		

be categorized than the former one. In Table 3 we may see the distribution of the two groups.

We have selected the K-means clustering method for the experiments carried out in this paper. 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 [7] 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 purpose of this paper, we have established the parameter k to be equal to two (companies and other).

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. (3)) with a similarity intra-tweets calculated by means of the cosine measure (see Eq. (4)).

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 (frequency), but is offset by the frequency of the term in the corpus. In this document, 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 [13] 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 normalized 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 normalized 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:

$$tf_{ij} = \frac{tf(t_i, d_j)}{\sum_{k=1}^{|d_j|} tf(t_k, d_j)} \quad (1)$$

$$idf(t_i) = log\left(\frac{|D|}{|d : t_i \in d, d \in D|}\right) \quad (2)$$

$$tf\text{-}idf = tf_{ij} * idf(t_i) \quad (3)$$

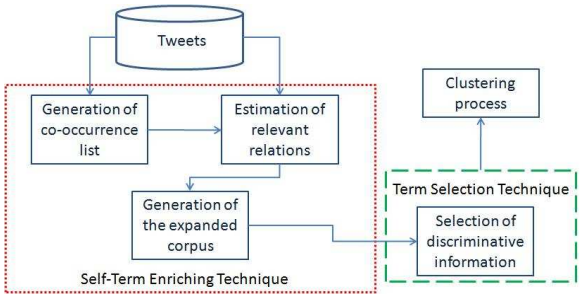

Figure 1: Self-Term Expansion Methodology

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, $\overrightarrow{d_i}$ and $\overrightarrow{d_j}$, it is possible to calculate the cosine of the angle between these two vectors as follows:

$$\text{Cos}_\theta(\overrightarrow{d_i}, \overrightarrow{d_j}) = \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{\left\|\overrightarrow{d_i}\right\| \left\|\overrightarrow{d_j}\right\|} \quad (4)$$

The similarity matrix is then constructed on the basis of the above formulae, i.e., for each possible pair of documents, 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.

We consider important to mention that the clustering algorithm (including the representation and matrix calculation) is applied after we have improved the representation of tweets in order to show the improvement of the enriching process.

The contribution of this paper is to propose and compare a number of different methods of representation of tweets based on term expansion and their impact on clustering company tweets. Therefore, there follows a detailed explanation of the different representation of the tweets that we have used for improving the clustering process.

4.1 Self-Term Expansion Methodology

The Self-Term Expansion Methodology (S-TEM) [11] 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., none 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.

4.1.1 Self-Term Enriching Technique.

The technique consists of replacing terms of a tweet with a set of co-related terms. A co-occurrence list is calculated from the target data set by applying Pointwise Mutual Information (PMI) [8]. PMI provides a degree of relationship

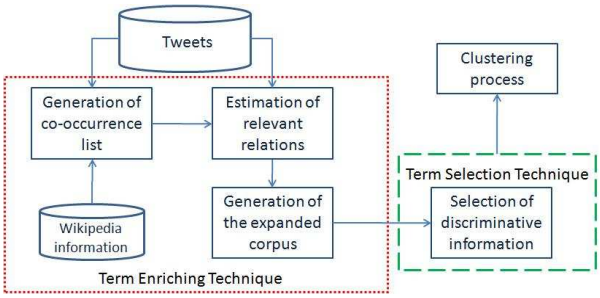

Figure 2: Term Expansion Methodology - Wiki

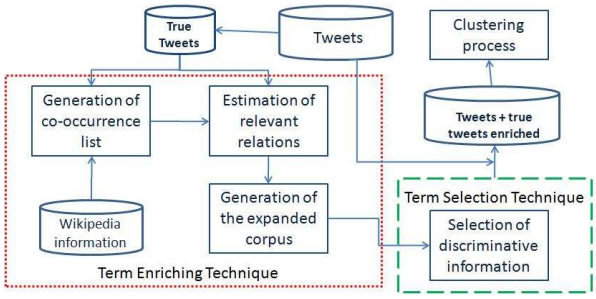

Figure 3: Term Expansion Methodology with Positive examples - Wiki

between two words; however, the level of this relationship must be empirically adjusted for each task. In this work, we established empirically the PMI value equal or greater than 2 to be the best threshold. In other experiments [11], a threshold of 6 was used; however in tweets correlated terms are rarely found because of the low term frequencies.

The Self-Term Enriching Technique is defined formally in [11] as follows: Let $D = \{d_1, d_2, ..., d_n\}$ be a document collection with vocabulary $V(D)$. Let us consider a subset of $V(D) \times V(D)$ of co-related terms as $RT = \{(t_i, t_j)|t_i, t_j V(D)\}$. The RT expansion of D is $D' = \{d'_1, d'_2, ..., d'_n\}$, such that for all $d_i \in D$, it satisfies two properties: 1) if $t_j \in d_i$ then $t_j \in d'_i$, and 2) if $t_j \in d_i$ then $t'_j \in d'_i$, with $(t_j, t'_j) \in RT$. If RT is calculated by using the same target data set, then we say that D' is the Self-Term Expansion version of D.

4.1.2 *Term Selection Technique.*

The Term Selection Technique helps us to identify the best features for the clustering process. However, it is also useful to reduce the computing time of the clustering algorithms. In particular, we have used Document Frequency (DF) [5], which assigns the value $DF(t)$ to each term t, where $DF(t)$ means the number of documents in a collection, where t occurs. The Document Frequency technique assumes that low frequency terms will rarely appear in other documents; therefore, they will not have significance on the prediction of the class of a document.

4.2 Term Expansion Methodology - Wiki (TEM-Wiki)

This technique is based on the two previously presented techniques (Self-Term Enriching Technique and the Term Selection Technique). The main difference of this technique with respect to the other two is that additionally of using a list of co-occurrence extracted from the same corpus, we have added valuable information in the enriching process which is extracted from Wikipedia[4]. Each company corpus was enriched with the corresponding company information provided by Wikipedia. In other words, we have extracted the information provided by Wikipedia of a particular company and we have enriched the particular company information before applying the clustering process. The aim of this approach is to improve the representation of tweets by using an external resource (in this case, Wikipedia). Wikipedia is an accessible and freely available resource that contains descriptions of the most recognized companies. Figure 2 illustrates the main components of this methodology.

[4]http://www.wikipedia.org/

4.3 Term Expansion Methodology with Positive examples - Wiki (TEM-Positive-Wiki)

This methodology proposes to improve the TEM-Wiki methodology by using exclusively positive samples, i.e., those tweets classified as "true". It uses the same information plus extra information related to the company, extracted from Wikipedia and information of positive samples in order to have a better criteria (information) to differentiate the company tweets related to a company from the non-related. We consider that companies are not transient entities or at least they have presence over Internet for a relatively long period of time. Accordingly we believe that enriching the *true* tweets can also help in for the clustering process i.e., the intention is to provide discriminatory information in order to ease up the identification of the groups or clusters. We consider this methodology slightly limited due to the use of the positive samples needed (in some cases, it is difficult to obtain samples of particular domains) but we also believe that it is necessary to be compared with the different approaches presented in this work. In Figure 3 we can see the diagram of the methodology.

4.4 Full Term Expansion Methodology (TEM-Full)

In this methodology we expand only the ambiguous word (the company name) with all the words that co-occur with it in the same class of the corpus without restrictions for the level of co-occurrence. 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 4. Notice that this expansion process does not use an external resource. We believe that due to the low term frequency and the shortness of the data (tweets), it is better to include all the information that co-occur in the corpus 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.

5. EXPERIMENTAL RESULTS

The aim of these experiments is to verify whether or not an enriching procedure would help for improving the task of clustering company tweets. Therefore, we have tested the four different methodologies proposed in the previous section over the two datasets constructed. In order to compare the performance of the different approaches, we have calculated one baseline which consists on clustering, with K-means, the tweets without any enriching procedure.

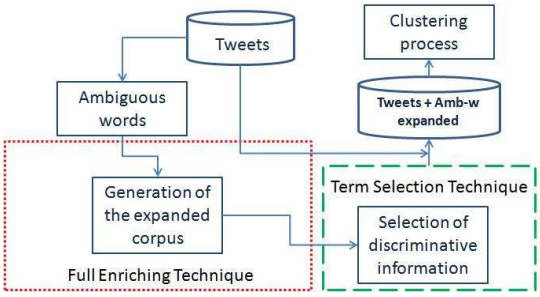

Figure 4: Full Term Expansion Methodology

The obtained results using the different methodologies proposed are compared in Table 4, the bold text represents the cases when the result is better than the baseline. We have compared the methodologies presented with the two subsets (generic and specific company names subsets) described previously.

In order to be objective with the results, we have used a well-known measure to evaluate the performance of the clustering algorithms, the F-measure [16], which is defined as follows: Given a set of clusters $C = \{C_1, ..., C_|c|\}$ and a set of classes $C^* = \{C_1^*, ..., C_|^*c^*|\}$, the F-measure (FM) between a cluster C_i and a class C_j^* is given in Equation (5):

$$FM(C_i, C_j^*) = \frac{2 * precision(C_i, C_j^*) * recall(C_i, C_j^*)}{precision(C_i, C_j^*) + recall(C_i, C_j^*)} \quad (5)$$

The global performance of a clustering method is computed by using F-measure values, the cardinality of the set of clusters obtained, and normalizing by the total number of documents $|D|$ in the collection. The obtained result is the F-measure and it is shown in Equation (6):

$$F\text{-Measure} = \sum_{1 \leq i \leq |C|} \frac{|C_i|}{|D|} max_{1 \leq j \leq |C^*|}(FM(C_i, C_j^*)) \quad (6)$$

We consider that there still some limitations on obtaining better F-measure results due to the particular writing style of tweets. There exists a poor grammatical structure and many out of vocabulary words, a fact that difficulties the task of clustering tweets. There is, however, a clear improvement by most of the approaches in comparison with the baseline. This is an indicative that the enriching procedure is a good technique of document representation. In particular, the term expansion technique increases the frequency of terms by adding co-related terms of the same class. This effect of enriching is shown to be effective for improving the clustering of company tweets.

Table 5 presents the number of times that each methodology has showed an improvement in the clustering process against the baseline. The TEM-Full methodology has shown the best performance with the corpus of generic company names.

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. We have observed that, no matter if we are using an external resource,

Table 5: A comparison of the different methodologies against the baseline

	Company names	
Methodologies	*Generic*	*Specific*
S-TEM	6	7
TEM-Wiki	6	9
TEM-Positive-Wiki	4	7
TEM-Full	10	7

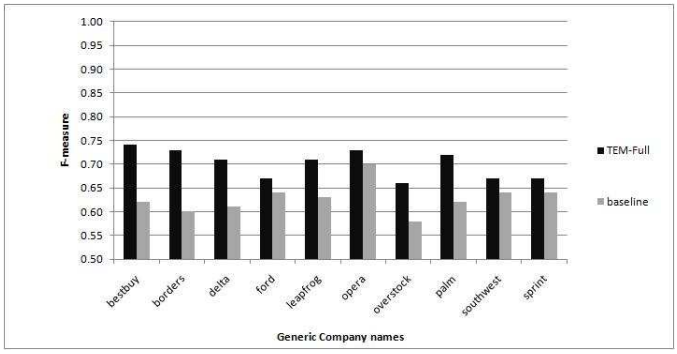

Figure 5: The TEM-Full methodology applied to the "generic" company name corpus

we may improve the representation of company tweets for clustering task.

In Figures 5 and 6 we show the performance of the two best 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 5). On the other hand, In Figure 6 it is possible to see that the TEM-Wiki methodology outperformed the baseline for the most of the company names.

Even if we have validated that the term expansion procedure is effective for improving the task of clustering company tweets, we are still interested in obtaining better F-Measure values than the ones we have obtained up to now. It is important to notice that the the F-measure values obtained by the TEM-Wiki and the TEM-Positive-Wiki approaches

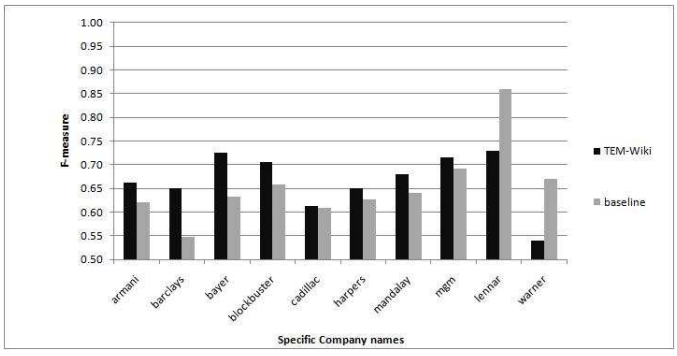

Figure 6: The TEM-Wiki methodology applied to the "specific" company name corpus

Table 4: 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).

Company	Methodologies				
	$S-TEM$	$TEM-Wiki$	$TEM-Positive-Wiki$	$TEM-Full$	*Baseline*
Generic Company Names Subset					
Bestbuy	**0.68**	**0.69**	0.62	**0.74**	0.62
Borders	**0.64**	**0.68**	0.60	**0.73**	0.60
Delta	**0.76**	**0.78**	0.61	**0.71**	0.61
Ford	0.60	0.60	**0.98**	**0.67**	0.64
Leapfrog	**0.69**	**0.69**	0.61	**0.71**	0.63
Opera	0.70	0.66	0.70	**0.73**	0.70
Overstock	**0.68**	**0.68**	**0.63**	**0.66**	0.58
Palm	**0.65**	**0.65**	0.60	**0.72**	0.62
Southwest	0.62	0.64	**1.0**	**0.67**	0.64
Sprint	0.60	0.60	**0.98**	**0.67**	0.64
Specific Company Names Subset					
Armani	**0.66**	**0.66**	**0.67**	**0.73**	0.62
Barclays	**0.65**	**0.65**	**0.65**	**0.72**	0.55
Bayer	**0.65**	**0.72**	**0.65**	**0.71**	0.63
Blockbuster	0.66	**0.71**	0.66	**0.71**	0.66
Cadillac	0.61	0.61	**0.64**	**0.69**	0.61
Harpers	**0.65**	**0.65**	**0.68**	**0.68**	0.63
Mandalay	**0.68**	**0.68**	**0.68**	**0.74**	0.64
Mgm	**0.72**	**0.72**	0.66	0.54	0.69
Lennar	**0.97**	**0.97**	**0.97**	0.72	0.96
Warner	0.61	**0.74**	0.67	0.54	0.67

are slightly different in comparison with the values obtained by the TEM-Full. This fact may lead us to consider that regardless of the resource used (internal or external), the clustering company tweets is a very difficult task.

6. CONCLUSIONS

Clustering short text corpora is a difficult task. Since tweets are also considered as short texts, the clustering task of tweets is also a complex problem to be solved. 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 leads to obtain low performances for most clustering methods.

In this paper we have presented different approaches for improving the task of clustering company tweets. In particular, 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, that 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 one. However, one of the four methodologies evaluated (TEM-Wiki) performed well on the former dataset and, another one methodology obtained the best results on the latter dataset (TEM-Full).

The two methodologies that obtained the best results are opposite in the sense of using or not external resources. TEM-Full is a completely unsupervised approach which constructs a thesaurus from the same dataset to be clustered and, thereafter, uses this resource for enriching the terms. On the other hand, TEM-Wiki uses information from Wikipedia that is introduced by human beings. We have used these different approaches in order to provide a better comparison and to have an idea 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) has shown slightly better results than the rest of the methodologies presented. The best results was obtained due to the inclusion of all the information that co-occur in the corpus of a particular company. It is important to say that this methodology (TEM-Full) showed a good performance due to the shortness and low term frequency of the data analyzed (tweets).

This is an initial effort for bringing order to Twitter by improving the term representation of this kind of texts. We expect to do further work in order to advance in the problem of clustering company tweets. In particular, we are interested in proposing highly scalable methods that may be able to deal with the huge amounts of information published every day in Twitter.

7. ACKNOWLEDGMENTS

This work has been partially supported by the projects: MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i), PROMEP #103.5/09/4213 and CONACYT #106625, as well as a grant provided by the Mexican Council of Science and Technology (CONACYT).

8. REFERENCES

[1] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proc. of the CICLing 2002 Conference*, pages 136–145. LNCS Springer-Verlag, 2002.

[2] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2007.

[3] A. Cheng and M. Evans. Inside twitter: An in-depth look inside the twitter world. Website, 2009. http://www.sysomos.com/insidetwitter/.

[4] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Ac, 1994.

[5] K. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[6] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD'03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge dicovery and data mining*, pages 137–146. ACM, 2003.

[7] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[8] D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[9] M. McGiboney. Twitter's tweets smell of success. Website, 2008. http://blog.nielsen.com/nielsenwire/online_mobile/twitters-tweet-smell-of-success.

[10] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas. *Twitter and the micro-messaging revolution: Communication, connections, and immediacy-140 characters at a time*. O'Really Report, 2008.

[11] D. Pinto. *On Clustering and Evaluation of Narrow Domain Short-Text Corpora*. PhD thesis, Universidad Politécnica de Valencia, 2008.

[12] Y. Qiu and H. Frei. Concept based query expansion. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169. ACM, 1993.

[13] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[14] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.

[15] B. Sriram, D. Fuhry, E. Demir, and H. Ferhatosmanoglu. Short text classification in twitter to improve information filtering. In *The 33rd ACM SIGIR'10 Conference*, pages 42–51. ACM, 2010.

[16] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.