On the Difficulty of Clustering Microblog Texts for Online Reputation Management

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

David Pinto
FCC, Benemérita Universidad
Autónoma de Puebla, Mexico
dpinto@cs.buap.mx

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

Paolo Rosso
NLE Lab. -ELiRF, Universidad
Politécnica de Valencia, Spain
prosso@dsic.upv.es

Abstract

In recent years microblogs have taken on an important role in the marketing sphere, in which they have been used for sharing opinions and/or experiences about a product or service. Companies and researchers have become interested in analysing the content generated over the most popular of these, the Twitter platform, to harvest information critical for their online reputation management (ORM). Critical to this task is the efficient and accurate identification of tweets which refer to a company distinguishing them from those which do not. The aim of this work is to present and compare two different approaches to achieve this. The obtained results are promising while at the same time highlighting the difficulty of this task.

1 Introduction

Twitter[1] - a microblog of the Web 2.0 genre that allows users to publish brief message updates - has become an important channel through which users can share their experiences or opinions about a product, service or company. In general, companies have taken advantage of this medium for developing marketing strategies.

Online reputation management - the monitoring of media and the detection and analysis of opinions about an entity - is becoming an important area of research as companies need up to the minute information on what is being send on the WWW about them and their products. Being unaware of negative comments regarding a company may affect its reputation and misguide consumers into not buying particular products. On the other hand companies may identify user feedback and use it in order to provide better products and services which could make them more competitive.

A first step in this process is the automatic collection of tweets relating to a company. In this paper we present an approach to the categorisation of tweets which contain a company name, into two clusters corresponding to those which refer to the company and those which do not. Clearly this is not as straightforward as matching keywords due to the potential for ambiguity. 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 (Milstein et al., 2008).

The rest of this paper is organised as follows. Section 2 describes the problem and the related work. Section 3 presents the data set used in the experiments. Section 4 explains the approaches 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 and Related Work

We are interested in discriminating between Twitter entries that correspond to a company from those that do not, 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 regard a company name as ambiguous if the word/s that comprise its name can be used in

[1] http://twitter.com

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 146–152,
24 June, 2011, Portland, Oregon, USA

different contexts. An example can be seen in Table 1 where the word $borders$ is used in the context of a company (row 1 & 3) and as the boundary of a country (row 2). We adapt a clustering approach to solving this problem although the size of tweets presents a considerable challenge. 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 two approaches 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 paper, can improve the representation of the microblogs from a clustering perspective, and as a consequence the performance of the clustering task. In addition, we test the hypothesis that specific company names - names that can not be found in a dictionary - such as $Lennar$ or $Warner$ may be more easily identified than generic company names such as $Borders$, $Palm$ or $Delta$, because of the ambiguity of the latter.

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

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
TRUE	33% Off Borders Coupon : http://wp.me/pKHuj-qj

We describe briefly here the work related to the problem of clustering short texts related to companies. In particular those works in the field of categorisation of tweets and clustering of short texts.

In (Sankaranarayanan et al., 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 et al., 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.

The main body of relevant related research emanates from the WePS-3 evaluation campaign in the task 2 called Online Reputation Management (Amigó et al., 2010). In (García-Cumbreras et al., 2010) the authors based their approach on recognising named entities, extracting external information and predefining rules. They use the well-known Name Entity Recogniser (NER) included in GATE[2] for recognising all the entities in their Tweets. They also use the web page of the organisation, Wikipedia and DBpedia[3]. Predefined rules are then applied to determine if a Twitter entry belongs to an organisation or not.

The work presented in (Kalmar, 2010) uses data from the company website. This data is used to create a initial model from which to bootstrap a model from the Tweets, the keywords and description are weighted. The features used are the co-occurring words in each tweet and the relevance of them was calculated according to the Pointwise Mutual Information value. Although it seems to be an interesting approach the results shown are disappointing.

In (Yerva et al., 2010) a support vector machine (SVM) classifier is used with the profiles built a priori. Profiles are constructed for each company which are sets of keywords that are related to the company or sets of keywords unrelated to the company. This system uses external resources such as Wordnet[4], meta-data from the company web page, GoogleSet[5] and user feedback. The research presented in (Yoshida et al., 2010) propose that organisation names can be classified as "organization-line names" or "general-word-like names". The authors have observed that the fact that ratio of positive or negative (if the tweet is related to the organisation or not) has a strong correlation with the types of organisation names i.e., "organization-like names" have high percentages of tweets related to the company

[2]http://gate.ac.uk/

[3]http://dbpedia.org/

[4]http://wordnet.princeton.edu/

[5]http://labs.google.com/sets

and when compared to "general-word-like names" Another approach is described in (Tsagkias and Balog, 2010), in which the authors trained the well-known J48 decision tree classifier using as features the company name, content value such as the presence of URLs, hashtags or is-part-of-a-conversation, content quality such as ratio of punctuation and capital characters and organisational context. This approach is quite interesting but they require a training set.

3 Dataset Description

We base our experiments on the corpus provided for task two of the WePS-3 evaluation campaign[6], related to Online Reputation Management for organisations, or specifically on the problem of organisation (company) name ambiguity.

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

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

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

The corpus was obtained from the *trial* and *training* data sets 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 corpus was labelled by five annotators: 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 where the annotators were unable to make a decision.

In order to gauge the problem and to establish a baseline for the potential of a clustering approach. We decided to cluster the data sets (trial and training) using the K-means algorithm (MacQueen, 1967) with k equal to three in order to have a clear reference and detect possible drawbacks that the collections may contain. The results were evaluated using the F-measure (van Rijsbergen, 1979) and gave values of 0.52 and 0.53 for the *trial* and *training* data sets respectively. This was expected, as clustering approaches typically work best with long documents and balanced groups (Perez-Tellez et al., 2009). Using this baseline, we then considered how a clustering approach could be improved by applying text enrichment methods. In order to compare only the effect of the enrichment however, we have modified the data set by including only those tweets written in English and for which a $true$ or $false$ label has been established, i.e., in the experiments carried out we do not consider 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 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 with 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. In the following section we present and compare the different approaches we propose for dealing with this problem.

4 Clustering Company Tweets

The purpose of this research work is to cluster tweets that contain a possible company entity into two

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

groups, those that refer to the company and those that refer to a different topic. We approach this problem by introducing and, thereafter, evaluating two 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. Although the idea has been previously studied in literature (Qiu and Frei, 1993; Grefenstette, 1994; Banerjee and Pedersen, 2002; Pinto et al., 2010) we are not aware of any work in which has applied it to microblog texts. In this paper, we evaluate the performance of two different approaches for term enriching in the task of clustering company tweets.

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 be clustered. The first group is composed of 10 companies with generic names, i.e., names that can be ambiguous (i.e., they have another common meaning and appear in a dictionary). 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 be categorised than the former one. In Table 3 we see the distribution of the two groups. We have selected the K-means clustering method (MacQueen, 1967) 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. 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 gained by applying the enriching process.

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		

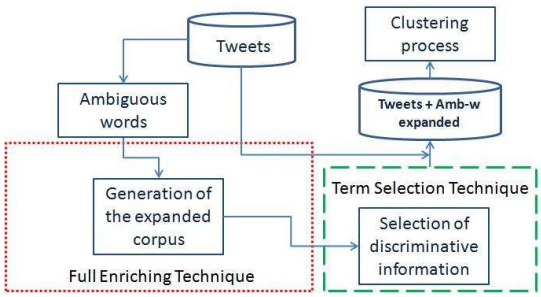

Figure 1: Full Term Expansion Methodology

4.1 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 alongside it, without restrictions for the level of co-occurrence. Our hypothesis states that the ambiguous words may bring important information from the identification of co-occurrence-relations to the next step of filtering relevant terms. 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 1. Note 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, it is better to include all the information that co-occurs in the corpus of a company and provide more information to the enriching process.

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.

4.2 Full Tem Expansion Methodology with a Text Formaliser (TEM-Full+F)

In this approach, we test the hypothesis that we can improve the cluster quality by increasing the level of formality in the document text. Due to the length restriction of 140 characters users tend to write comments using abbreviations. We have used an abbreviation dictionary[7] that contains 5,173 abbreviations commonly used in microblogs, tweets and short messages. After the formalisation step, the expansion is performed but it is only applied to the ambiguous word (the company name) and words

[7] http://noslang.com/dictionary

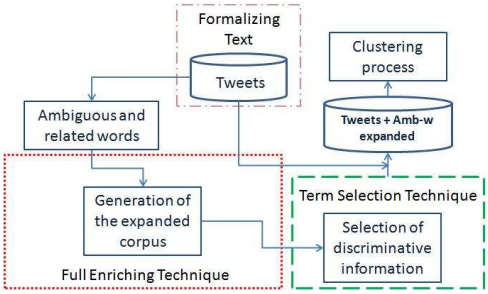

Figure 2: Full Term Expansion Methodology with a Text Formaliser (TEM-Full+F)

which highly co-occur with it. These words were selected as they appear in frequently with the ambiguous word in positive tweets (i.e., those related to the companies). We consider that this kind of word may help us take the correct decision during the clustering process because they are highly related with the company tweets. The words selected to be expanded were closely related to the company such as crew, jet, flight, airlines, airplane for *Delta* company name. In the case of the *Opera* company name the words expanded were software, technology, developers, interface, web, browser. The number of words per company name were between five and ten, showing that even a small number of words that co-occur highly may help in the enriching process. We have used the Term Selection Technique as described in 4.1 and no external resource. The process is shown in Figure 2.

5 Experimental Results

In this section we present the results obtain by the related approaches and also the results obtained by our methodologies proposed.

5.1 Related Approaches

Although the results are not directly comparable with our approaches due to the slightly different dataset used in the experiments (see Section 3), we would like to provide a clear description of the different approaches with the objective of highlight the strengths of the related approaches developed for this purpose.

In Table 4, the best results (F-measure related classes) reported by the approaches presented to the task two of the WePS-3 evaluation campaign

Table 4: Related approaches (F-measure related)

Approaches				
L	S	I	U	K
0.74	0.36	0.51	0.36	0.47

L = LSIR-EPFL, S = SINAI, I = ITC-UT, U = UVA, K = KALMAR

(Amigó et al., 2010). It is important to mention that all these systems used the whole collection even if the companies subsets where very imbalanced. In our case, we are interested in proposing approaches that can deal with two different kind of company names such as “generic” and “specific” rather than one methodology for both.

In Table 4 the LSIR-EPFL system (Yerva et al., 2010) showed very good performance even when the subsets are very imbalanced. The SINAI system (García-Cumbreras et al., 2010) took advantage of the entity recognition process and they report that named entities contained in the microblog documents seem to be appropriate for certain company names. ITC-UT (Yoshida et al., 2010) incorporated a classifier and made use of Named Entity Recognition and Part-of-Speech tagger is also good in their performance but as the authors in (Amigó et al., 2010) have mentioned “it is difficult to know what aspect lead the system to get ahead other systems” as each takes advantage of different aspects available such as external resources or tools. UVA (Tsagkias and Balog, 2010) is an interesting contribution but the only problem is training data will not always be available for some domains. Finally, the KALMAR system (Kalmar, 2010) seems to achieve good performance when applied to well-balanced collections. In contrast to these approaches, we would like to emphasize that our approaches are predominantly based on the information to be clustered.

5.2 Results of Our Experiments

In order to present the performance of the different proposed approaches, we have calculated a baseline based on clustering, with K-means, and with no enriching procedure. The obtained results using the two methodologies are compared in Table 5. We have shown in bold text the cases in which the result equalled or improved upon the baseline. We have compared the methodologies presented with the two

subsets (generic and specific company names subsets) described previously.

Table 5: A comparison of each methodology with respect to one baseline using the F-measure.

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

B - Baseline, OA - Overall Average

We consider that there still some limitations on obtaining improved results due to the particular writing style of tweets. The corpus exhibits a poor grammatical structure and many out-of-vocabulary words, a fact that makes the task of clustering tweets very difficult. There is, however, a clear improvement in most cases in comparison with the baseline. This indicates that the enriching procedure yields benefits for the clustering process.

The TEM-Full methodology has demonstrated good performance with the corpus of generic company names with 0.70 average (F-measure value) 8 points over the average baseline. In this case, we have expanded only the ambiguous word (the name of the company), whereas the TEM-Full+F methodologies performed well (0.74 F-measure) with the corpus of specific company names. We have observed that, regardless of whether or not we are using an external resource in TEM-Full and TEM-Full+F approaches, we may improve the representation of company tweets for the clustering task. It is important to mention that the good results presented in companies such as $Bestbuy$ or $Lennar$ were obtained because the low overlapping vocabulary between the two categories (positive and negative) and, therefore, the clustering process could find well-delimited groups. We also would like to note that sometimes the methodologies have produced only minor performance improvement. This we believe is largely due to the length of the tweets, as it has been demonstrated in other experiments that better results can be achieved with longer documents (Perez-Tellez et al., 2009; Pinto et al., 2010).

The best result has been achieved with the TEM-Full+F methodology which achieved an overall average F-measure value 0.72, it is 8 points more than the overall average of the baseline. This methodology has not disimproved on the baseline 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 by detecting and filtering stronger relations that may help in the identification of the positive company tweets. 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 by definition short texts (having a maximum of 140 characters), the clustering of tweets is also a challenging problem as stronger results typically achieved with longer text documents. Furthermore, due to the nature of writing style of these kinds of texts - typically they exhibits an informal writing style, with poor grammatical structure and 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 two different approaches for representing tweets on the basis term expansion and their impact on the problem of clustering company

tweets. In particular, we introduced two 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 data set and the enriching methodology is not applied.

We consider that TEM-Full performed well on the former data set and, another methodology obtained the best results on the latter data set TEM-Full+F. However, the TEM-Full+F methodology appears suitable for both kinds of corpora, and does not require any external resource. TEM-Full and TEM-Full+F are completely unsupervised approaches which construct a thesaurus from the same data set to be clustered and, thereafter, uses this resource for enriching the terms. On the basis of the results presented, we can say that using this particular data, the unsupervised methodology TEM-Full+F has shown improved results.

This paper has reported on our efforts to apply clustering and term enrichment to the important problem of company identification in microblogs. We expect to do further work in proposing highly scalable methods that may be able to deal with the huge amounts of information published every day in Twitter.

Acknowledgments

This work was carried out in the framework of the MICINN Text-Enterprise TIN2009-13391-C04-03 research project and the Microcluster VLC/Campus (International Campus of Excellence) on Multimodal Intelligent Systems, PROMEP #103.5/09/4213 and CONACYT #106625, as well as a grant provided by the Mexican Council of Science and Technology (CONACYT).

References

E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. 2010. WePS-3 evaluation campaign: Overview of the online reputation management task. In *CLEF 2010 (Notebook Papers/LABs/Workshops)*.

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

M. A. García-Cumbreras, M. García Vega, F. Martínez Santiago, and J. M. Perea-Ortega. 2010. Sinai at weps-3: Online reputation management. In *CLEF 2010 (Notebook Papers/LABs/Workshops)*.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Ac.

P. Kalmar. 2010. Bootstrapping websites for classification of organization names on twitter. In *CLEF 2010 (Notebook Papers/LABs/Workshops)*.

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

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

F. Perez-Tellez, D. Pinto, Cardiff J., and P. Rosso. 2009. Improving the clustering of blogosphere with a self-term enriching technique. In *Proc. of the 12th Int. Conf. on Text, Speech and Dialogue*, pages 40–49. LNAI.

D. Pinto, P. Rosso, and H. Jimenez. 2010. A self-enriching methodology for clustering narrow domain short texts. *The Computer Journal*, doi:10.1093/comjnl/bxq069.

Y. Qiu and H.P. Frei. 1993. Concept based query expansion. In *Proc. of the 16th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 160–169. ACM.

J. Sankaranarayanan, H. Samet, B.E. Teitler, M.D. Lieberman, and J. Sperling. 2009. Twitterstand: news in tweets. In *Proc. of the 17th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, pages 42–51. ACM.

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

M. Tsagkias and K. Balog. 2010. The university of amsterdam at weps3. In *CLEF 2010 (Notebook Papers/LABs/Workshops)*.

C.J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

S. R. Yerva, Z. Miklós, and K. Aberer. 2010. It was easy, when apples and blackberries were only fruits. In *CLEF 2010 (Notebook Papers/LABs/Workshops)*.

M. Yoshida, S. Matsushima, S. Ono, I. Sato, and H. Nakagawa. 2010. Itc-ut: Tweet categorization by query categorization for on-line reputation management. In *CLEF (Notebook Papers/LABs/Workshops)*.