Computational Collective Intelligence. Technologies and Applications: 5th International Conference, ICCCI 2013, Craiova, Romania, September 11-13, 2013, Proceedings. Lecture Notes in Computer Science, Volumen 8083. Springer, 2013. 622-630.

DOI: http://dx.doi.org/10.1007/978-3-642-40495-5_62

Extracting Collective Trends from Twitter using Social-based Data Mining

Gema Bello[1], Héctor Menéndez[1], Shintaro Okazaki[2], and David Camacho[1]

[1]Departamento de Ingeniería Informática. Escuela Politécnica Superior. Universidad Autónoma de Madrid.
C/Francisco Tomás y Valiente 11, 28049 Madrid, Spain
[2]Department of Finance and Marketing Research. College of Economics and Business Administration. Universidad Autónoma de Madrid.
C/Francisco Tomás y Valiente 5, 28049 Madrid, Spain
{gema.bello,hector.menendez,shintaro.okazaki,david.camacho}@uam.es

Abstract. Social Networks have become an important environment for Collective Trends extraction. The interactions amongst users provide information of their preferences and relationships. This information can be used to measure the influence of ideas, or opinions, and how they are spread within the Network. Currently, one of the most relevant and popular Social Network is Twitter. This Social Network was created to share comments and opinions. The information provided by users is specially useful in different fields and research areas such as marketing. This data is presented as short text strings containing different ideas expressed by real people. With this representation, different Data Mining and Text Mining techniques (such as classification and clustering) might be used for knowledge extraction trying to distinguish the meaning of the opinions. This work is focused on the analysis about how these techniques can interpret these opinions within the Social Network using information related to IKEA® company.

Keywords: Collective Trends, Social Network, Data Mining, Classification, Clustering, Twitter

1 Introduction

Data Mining techniques have become an important field with several applications over the last few years[13]. Some of these applications have been oriented to Social Networks which contain a lot of information about their users, specially preferences, opinions and ideas[3]. Using this data, different companies have focused their marketing strategies on the influence of their products in their potential clients[3].

Currently, one of the most popular Social Networks is Twitter [1]. This Network allows its users to communicate between them using text string of 140 characters. It becomes a Collective Intelligence where the users generate an emergency information source through their comments about different topics. Twitter has several APIs to extract the information provided by the users, which offer new research challenge in different science fields[18].

Document clustering techniques can be applied for efficient organization, navigation, retrieval, and summary of huge volumes of text documents [19, 9, 15]. These methods can automatically organize a document corpus into clusters or similar groups which allow the knowledge extraction about user behaviour. The clustering techniques were designed to find hidden information or patterns in a dataset. They are based on a blind search in an unlabelled data collection, grouping the data with similar properties in clusters without the necessity of labelled data or human supervision. The topic detection problem can be considered as a special case of the document clustering, therefore, these techniques can be used over the textual messages provided by Twitter to extract the conversation topics and then detect collective trends from the data.

This work has been oriented on the identification of the types of comments which are provided from the users about the quality of a concrete company, in this case IKEA® . The present method can be applied to understand Twitter sentiment trends regarding companies, extracting the community mood based on a small set of tweets gathered at an instant of time. It shows how different classification and clustering techniques can be used to extract this information from the Social Networks. Finally, a comparative study of these techniques that have been applied to message text collected is presented.

The rest of the paper is structured as follows: Section 2 shows the Related Work and presents the classification and clustering techniques used during the analysis. Section 3 explains the metrics used for the model validation phase of the analytical process. Section 4 is focused on the experiments which have been carried out and their results. Finally, the last section presents the Conclusions and Future Work.

2 Related Work

Techniques of Collective Intelligence have been used in several fields. These techniques are based on the intelligence emerged by the groups which compete or collaborate in an environment. It has applications to Biology, Psychology and Computer Networks, amongst others. This work is focused on trends extraction, similar to [3] where Data Mining techniques are applied to extract information of users from electronic commerce. This information is related to ideas, preferences and behaviours of the users and the interest of the users where they are trying to find products according to similar users preferences and opinions.

Data Mining classical techniques have been applied in the Collective Trends extraction process. Different classification and clustering methods have been compared trying to find the best approach. Following subsections introduce the techniques used in this work.

2.1 Classification Techniques

The data classification techniques which have been used are the following:

- **C4.5 trees**: C4.5 [17] technique is the most classical technique in data classification. It divides the data linearly using limits in the attributes and generates a decision tree. The division is chosen using a metric such as the data entropy.
- **Naive Bayes**: The Naive Bayes (NB) [8] classifier considers each feature independent to the rest of the features. Each of them contributes to the model information. It is based on Bayes Probability Laws.
- **K-Nearest Neighbours**: K-Nearest Neighbour algorithm (KNN) [6] classifies an element according to its neighbours. Depending on the K value, it considers the K-nearest neighbours and estimate the value of the data instance which is not classified.
- **Support Vector Machines**: Support Vector Machines (SVM) [5] usually changes the dimension of the search space through different kernel functions trying to improve the classification through a hyperplane separation of the data instances in the expanded space.

2.2 Clustering Techniques

Document clustering has been studied intensively because of its wide application in areas such as Web Mining [19], Search Engine and Information Retrieval [9, 15]. This technique allows the automatic organization of documents into clusters or groups [7]. Documents within a cluster have high similarity in comparison to one another, but are very dissimilar to documents in other clusters [12]. The grouping is based on the principle of maximizing intra-cluster similarity and minimizing inter-cluster similarity [2, 14].

In this paper K-Means which is a partitioning clustering algorithm, is applied to obtained the clusters or topics of the Tweets extracted from Twitter. It is a simple and well known algorithm for clustering [11]. All items are represented as a set of numerical features, and the number of resulting clusters (k) must be fixed before the algorithm has been executed. Then the algorithm randomly chooses k points in vector space such as the initial cluster centers. Afterwards, each item is assigned to the closer center using the distance measure chosen. After that, for each cluster a new center is calculated by averaging the vectors of all items assigned to it. The process of assigning items and recalculate centers is repeated until the process converges or a number of iterations is completed.

3 Model Validation Metrics

The validation metrics which have been used to measure the quality of the classification algorithms are Precision, Recall and F-Measure. These metrics are defined as follow [16]:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Where tp represents true-positives, fp represents false-positives and fn represents false-negatives.

Precision is used to measure the situation when an instance which does not belong to the class set is classified as part of the class set. **Recall** measures the situation when an instance is rightly classified according to its class. The **F-measure** is a metric which balances these measures.

4 Experiments

This section describes the experiments carried out in this work. The first part describes the data extraction process. The second part explain the data preprocessing processes which have been used during the analysis. The third part shows the experimental setup. Finally, the last part is focused on the results obtained and their interpretation.

4.1 Data Extraction

The data which have been analysed in this work comes from Twitter. Twitter is a Social Network where people usually publish information about personal opinions. It is divided in two kind of users behaviour: follower and following. As a follower, the user receives information of people which is followed by him, and as a following, the user information is sent to its followers. The information that the users share is called Tweets. Tweets are sentences limited by 140 characters which can contain information about personal opinions of the users, photos, links, etc. A user can also re-tweet the information of other users and share it.

The information extracted for the analysis are 100 comments about IKEA®. The comments have been extracted from '11-02-2013 15:24' to '18-02-2013 15:25', all comments come from different users (there are 100 users), the comments have been taken from Spain and the language is Spanish. These comments have been classified by marketing experts in four categories:

1. **Exclusion**: Those comments which are provided by companies to advertise their products. The class corresponds with the 8% of the total tweets.
2. **Satisfaction**: Positive information of the users about a product. The class corresponds with the 31% of the total tweets.
3. **Dissatisfaction**: Negative information of the users about a product. The class corresponds with the 29% of the total tweets.
4. **Neutral**: Neutral information of the users about a product. The class corresponds with the 37% of the total tweets.

4.2 Data Preprocessing

Due to the different techniques need different preprocessing, two methods have been used according to the nature of the process applied: classification and clustering.

Data Preprocessing for Classification The Preprocessing process consists in some typical steps oriented to simplified the text information. In this case, the preprocessing has been divided in three steps:

1. Eliminate Stop-Words and special characters of the sentences.
2. Generate a term-document matrix with the keywords.
3. Use a feature selection technique to choose the most relevant words for the analysis and reduce the search space.

The original term-document matrix is formed by 747 attributes. The Feature Selection technique used is the Correlation-based Feature Subset Selection [10] combined with an Exhaustive Search. The final term-matrix has 15 attributes: 'bien', 'millones', 'todo', '#publicidad ', 'bonita', 'estás', 'hacer', 'pues', 'quiero', 'toca', 'has', 'llevo', 'más', 'saben', 'solo'.

Data Preprocessing for Clustering A usual model for representing the content of document or text is the vector space model. In this vector space model each document is represented by a vector of frequencies of remaining terms within the document [9]. The term frequency (TF) is a function of the number of occurrences of the particular word in the document divided by the number of words in the entire document. Other function usually use is the inverse document frequency(IDF), typically, documents are represented as a TF-IDF feature vectors. With this data representation a document represents a data point in d-dimensional space where d is the size of the corpus vocabulary.

Text documents are tokenized transforming them in TF-IDF vectors. This step has included stop words removal and stemming on the document set. Besides a log normalization is applied to cleaning up edge data cases and then the TF-IDF vectors are generated which are used for further clustering process.

4.3 Experimental Setup

The experiments have been carried out using the classification and clustering algorithms describes in Sections 2.1 and 2.2. The parameters and metrics selection can be found in Table 1.

Algorithm	Parameters	Metric
Naive Bayes	-	-
C4.5	Confidence factor = 0.25 Min. Number objects = 2	Information Entropy
SVM	$\sigma = 0.1$	RBF[1]
K-Nearest Neighbour	$K = 5$	Euclidean Distance
K-Means	$K = 3 \ldots 5$	Euclidean Distance

Table 1. Parameters and metrics selection for the techniques.

All the classification algorithms have been validated using a 10-Cross Fold validation process.

4.4 Experimental Results

Table 2 shows the results of the analysis applying the classification techniques defined above. The metrics used are Precision, Recall and F-Measure.

The first technique, NB, obtains the best results for the classification according to the F-measure metric. It has problems to classify the first class (Exclusion) which are those comments introduced by companies, however the rest of the techniques obtain worse results. Satisfaction and Dissatisfaction obtains generally good results for NB, although SVM achieves similar results. For the Neutral class all the algorithms have good results, therefore this class is easier to identified for classifiers. The highest Precision value is achieved by C4.5 for Dissatisfaction and the highest Recall value is achieved by SVM and NB for Neutral. Since the best F-measure values (the balanced metric of Precision and Recall) for the classes are achieved by NB, it is considered the best classifier of this analysis.

There are some details which also should be mentioned related to the classification analysis:

- The Exclusion class is difficult to distinguish in almost all the cases.
- The Neutral class is clearer separated from the others.

The clustering results have shown that the best K-value for the K-means algorithm is 5. These results are concluded from both the number of discriminate classes and the F-measure metric. For each K-value: 3-means and 4-means only discriminate two classes from the original analysis (Neutral and Satisfaction) while 5-means also separates Dissatisfaction. The F-measure shows that

[1] Radial Basis Function [4]: this metric is defined by $e^{-\sigma||u-v||^2}$

Technique	Class	Cluster Num	Precision	Recall	F-Measure
NB	Exclusion	-	0.5	0.25	**0.333**
	Neutral	-	0.61	0.973	**0.75**
	Satisfaction	-	0.652	0.484	**0.556**
	Dissatisfaction	-	0.692	0.391	**0.556**
KNN	Exclusion	-	0.2	0.25	0.222
	Neutral	-	0.605	0.703	0.65
	Satisfaction	-	0.5	0.387	0.436
	Dissatisfaction	-	0.5	0.478	0.489
C4.5	Exclusion	-	0	0	0
	Neutral	-	0.45	0.973	0.615
	Satisfaction	-	0.714	0.323	0.444
	Dissatisfaction	-	1	0.174	0.296
SVM	Exclusion	-	0	0	0
	Neutral	-	0.621	0.973	0.758
	Satisfaction	-	0.571	0.516	0.542
	Dissatisfaction	-	0.769	0.435	0.556
3-Means	Exclusion	0	-	-	-
	Neutral	1	0.385	0.921	0.543
	Satisfaction	2	0.667	0.100	0.174
	Dissatisfaction	0	-	-	-
4-Means	Exclusion	0	-	-	-
	Neutral	2	0.387	0.780	0.491
	Satisfaction	2	0.714	0.0876	0.154
	Dissatisfaction	0	-	-	-
5-Means	Exclusion	0	-	-	-
	Neutral	2	0.402	0.802	**0.509**
	Satisfaction	2	0.833	0.113	**0.196**
	Dissatisfaction	1	0.5	0.0435	**0.080**

Table 2. Results of the application of the different models using Precision, Recall and F-measure metrics for validation.

3-means has the best value for Neutral class and 5-means has the best value for Satisfaction, however, both F-measure value results are closed using these algorithms. Hence, since 5-means can distinguish the Dissatisfaction class, it has been chosen as the best clustering results.

Analysing the number of clusters related to each class (in 5-means results), there are several aspects which are remarkable:

- The Neutral and Satisfaction classes, which are the most predominant, can be separated in two sub-trends per class. It means that a more detailed analysis of these trends would perform a better separation of the users opinions.
- The Exclusion class is undistinguishable in all cases. It means that this class should not be considered as a trend in the Tweets.

Comparing classification an clustering techniques, there are several things that are concluded: classification techniques obtains better results than clustering techniques, it is a consequence of the nature of the methods, clustering is a blind process while classification is a supervised process. However, applying the clustering techniques, a higher number of trends is obtained which allows a more detailed analysis of the conversations. Also clustering does not need a previous human-labelling process which is really problematic for huge datasets.

The clustering techniques have similar results distinguishing the Neutral class (which is the predominant class) as the classification methods. Also, they are not able to distinguish the Exclusion class. Hence, Exclusion should not be considered as a trend.

5 Conclusions and Future Work

This work has shown the application of Data Mining methods to extract Collective Trends from Twitter. A human-labelled dataset, extracted from Tweets of different users about IKEA®, has been used for the analysis. Clustering and classification techniques have been applied to extract the trends of users opinions and also to compare their results.

The different techniques have proved to be useful for this kind of analysis. However, these techniques are not enough to distinguish the classes. Classification techniques have achieved better results than clustering techniques, however, clustering techniques do not need to have the predefined classes for their application which is more useful for larger datasets. In addition, clustering techniques also provides more detailed information about the trends. It suggests that a clustering technique should be helpful for the initial human-labelling process.

Future work will be focused on the combination of both, classification and clustering techniques, to improve the trends identification using a previous clustering process to guide the human-labelling work. In addition, a more complete clustering study might be applied using more complex techniques to make a deeper trend study.

Acknowledgments

The preparation of this manuscript has been supported by the Spanish Ministry of Science and Innovation under the following projects: TIN2010-19872, ECO2011-30105 (National Plan for Research, Development and Innovation) and the Multidisciplinary Project of Universidad Autónoma de Madrid (CEMU-2012-034). The authors thank Ana Daz and Mercedes Rozano for the manual classification of the comments.

References

1. Twitter web site, 2013. twitter.com.
2. H. Ahonen-Myka. Mining all maximal frequent word sequences in a set of sentences. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 255–256, New York, NY, USA, 2005. ACM.
3. T. Bruckhaus. Collective intelligence in marketing. In J. Casillas and F. Martnez-Lpez, editors, *Marketing Intelligent Systems Using Soft Computing*, volume 258 of *Studies in Fuzziness and Soft Computing*, pages 131–154. Springer Berlin Heidelberg, 2010.
4. M. D. Buhmann and M. D. Buhmann. *Radial Basis Functions.* Cambridge University Press, New York, NY, USA, 2003.
5. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
6. T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21 –27, january 1967.
7. D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 318–329, New York, NY, USA, 1992. ACM.
8. P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, 29(2-3):103–130, Nov. 1997.
9. W. B. Frakes and R. A. Baeza-Yates, editors. *Information Retrieval: Data Structures & Algorithms.* Prentice-Hall, 1992.
10. M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.
11. J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.
12. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley-Interscience, 9th edition, Mar. 1990.
13. D. T. Larose. *Discovering Knowledge in Data.* John Wiley and Sons, 2005.
14. Y. Li, S. M. Chung, and J. D. Holt. Text document clustering based on frequent word meaning sequences. *Data Knowl. Eng.*, 64(1):381–404, Jan. 2008.

15. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.
16. D. M. W. Powers. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia, 2007.
17. J. R. Quinlan. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
18. D. N. Trung, J. J. Jung, N. Lee, and J. Kim. Thematic analysis by discovering diffusion patterns in social media: an exploratory study with tweetscope. In *Proceedings of the 5th Asian conference on Intelligent Information and Database Systems - Volume Part II*, ACIIDS'13, pages 266–274, Berlin, Heidelberg, 2013. Springer-Verlag.
19. O. Zamir and O. Etzioni. Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA, 1998. ACM.