COMBINING SOCIAL-BASED DATA MINING TECHNIQUES TO EXTRACT COLLECTIVE TRENDS FROM TWITTER

Gema Bello-Orgaz[1], Héctor Menéndez[2], Shintaro Okazaki[3] and David Camacho[4]
[1, 2, 4] Department of Computer Science,
[3] Department of Marketing,
Universidad Autónoma de Madrid
28049 – Madrid, Spain

[1]gema.bello@uam.es, [2]hector.menendez@uam.es, [3]shintaro.okazaki@uam.es, [4]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 Networks is Twitter. This Social Network was created to share comments and opinions. The information provided by users is especially 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 techniques (such as classification or clustering) will be used for knowledge extraction to distinguish the meaning of the opinions. Complex Network techniques are also helpful to discover influential actors and study the information propagation inside the Social Network. This work is focused on how clustering and classification techniques can be combined to extract collective knowledge from Twitter. In an initial phase, clustering techniques are applied to extract the main topics from the user opinions. Later, the collective knowledge extracted is used to relabel the dataset according to the clusters obtained to improve the classification results. Finally, these results are compared against a dataset which has been manually labelled by human experts to analyse the accuracy of the proposed method.

Keywords: Collective Trends, Social Network, Classification, Clustering, Twitter.

1. INTRODUCTION

Data Mining techniques have become an important field in Computer Science and Engineering with several applications over the last few years [18]. Some of these applications have been oriented to Social Networks which store a huge amount of information about their users, specially related to their preferences, opinions and ideas [28]. Using this data, different companies have focused their marketing strategies on the influence of their products in their potential clients [5].

Complex Network techniques have been also used to study Social Networks and their influence in Marketing [10]. These methods can be used to detect and analyse different aspects about the network such as its structure, the strength of its connections, communities formed by the interactions, etc., which are useful to understand how the users interact inside the network [15].

Currently, one of the most popular Social Networks is Twitter [29]. This Network allows its users to communicate between them using text string of 140 characters. It becomes a Collective Knowledge and Trend network where the users generate a direct (and summarized) information source through their comments about different topics. Twitter implements several APIs to automatically extract the information provided by the users, which offer a new research challenge in different science fields, such as Text Mining, Semantics, Data Mining and Information Retrieval amongst others [14, 28].

The first part of the work is focused on Text Mining methods. These techniques can be applied for efficient organization, navigation, retrieval, and summary of huge volumes of text documents [13, 22, 32]. These methods can automatically organize a document corpus into similar groups which allow the knowledge extraction about user behaviour, opinions or trends. Classification and clustering techniques are the most common Text Mining methods. Clustering techniques are based on a blind search in an unlabelled data collection while Classification techniques used labelled data to define the patterns.

This part combines clustering and classification for sentiment analysis of a labelled tweet dataset which contains user opinions. Firstly, clustering techniques are applied to extract the main topics and generate the topic clusters. The topic detection problem can be considered as a special case of the document clustering problem. 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. Using the information obtained from these clustering methods, the tweet dataset which has been previously classified by humans is relabelled according to the new clusters generated. Finally, a comparative study with our previous work [3] is presented. Classification techniques are applied to compare the results obtained according to the previous work, and the results obtained with the new class discrimination, improved through an initial clustering analysis, are shown.

The second part of the work is focused on the Tweets Network structure. First, the different communities generated by the re-tweets and mentions inside the messages are typified through a graph representation, and later PageRank is applied to find the most relevant actors of these communities. The information provided by this part of the analysis is related to the influential actors, and how the different communities are constructed.

This second analysis introduces two different perspectives about Social Data Mining, one based on Data Mining knowledge extraction, and the second one related to the utilization of PageRank to retrieve relevant features from the network structure. In our approach, we take advantage of the comments which are provided from the users about the quality of a concrete company, in this case IKEA®. The methodology presented in this work can be applied to understand Twitter sentiment trends regarding companies, and to extract the community mood based on a small set of tweets gathered at an instant of time. Finally, this work shows how different techniques can be used to extract this collective knowledge information from Social Networks.

The rest of the paper is structured as follows: Section 2 shows the Related Work and presents the classification, clustering and complex network techniques used during the analysis. Section 3 describes the different phases of the methodology applied. Section 4 explains the experimental setup used and the experimental results. Finally, the last section presents the main conclusions of this work.

2. RELATED WORK

Data Mining techniques have been used in several fields such as Biology, Psychology, Marketing and Computer Networks, amongst others [5, 10, 21]. These techniques are used to extract knowledge based on the intelligence emerged by the groups which compete or collaborate in an environment [3, 15, 28]. This work is focused on trends extraction and propagation, similarly to [4], 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 their interests when they are trying to find products according to similar user preferences and opinions.

In this work, clustering and classification techniques are combined to extract collective knowledge from Twitter [29]. Firstly, clustering techniques are applied to extract the main topics from the user opinions. Later, the collective knowledge extracted is used to relabel the dataset according to the clusters obtained, in order to improve the classification results. The following subsections introduce briefly all the techniques and algorithms used in this new approach.

2.1 Classification Techniques

Classification techniques have been widely used in Data Mining [18]. These techniques consist of the process to find patterns in data supervising the search through labels which define the instance categories. There are several classical techniques in the literature which have been applied and studied in different domains [6, 7, 9, 18, 25]. In this work, the data classification techniques which have been used are the following:

- **C4.5 trees**: The C4.5 technique [25] is one of the most classical ones 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) classifier [9] considers each feature independent of 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) [7] classifies an element according to its neighbours. Depending on the K value, it considers the K-nearest neighbours and estimates the value of the data instance which is not classified.
- **Support Vector Machines**: Support Vector Machines (SVM) [6] usually changes the dimension of the search space through different kernel functions trying to improve the classification through a hyper plane separation of the data instances in the expanded space.

2.2 Clustering Techniques

Document clustering techniques have been studied intensively because of their wide application in areas such as Web Mining [32], Search Engines [4] and Information Retrieval [13, 22]. These techniques allow the automatic organization of documents into clusters or groups [8]. Documents within a cluster have high similarity among them, but are very dissimilar to other documents in different clusters [17]. The documents (or items) grouping is based on the principle of maximizing intra-cluster similarity and minimizing inter-cluster similarity [1, 20].

In this paper, K-Means which is a partitioning clustering algorithm, is applied to obtain the clusters or topics of the Tweets extracted from Twitter. It is a simple and well known algorithm for clustering [16]. All items are represented as a set of numerical features, and the number of resulting clusters (*k*) must be fixed before the algorithm can be executed. Then the algorithm randomly chooses *k* points in vector space such as the initial cluster centres. Afterwards, each item is assigned to the closer centre using the distance measure chosen. After that, for each cluster, a new centre is calculated by averaging the vectors of all items assigned to it. The process of assigning items and recalculate centres is repeated until the process converges, or a number of iterations are completed.

The clustering algorithms which have been applied in this work are the following:

- **K-means Algorithm:** It is a simple and well known algorithm for clustering [16]. All items are represented as a set of numerical features (in this case TF-IDF vectors), and the number of resulting clusters (*k*) must be fixed before the algorithm can be executed. Then the algorithm randomly chooses *k* points in vector space such as the initial cluster centres. Afterwards, each item is assigned to the closer centre using the distance measure chosen. After that, for each cluster, a new centre is calculated by averaging the vectors of all items assigned to it. The process of assigning items and recalculate centres is repeated until the process converges or a number of iterations are completed. The algorithm can be proven to converge after a finite number of iterations.

- **Fuzzy K-means Algorithm:** This algorithm is an extension of K-Means. Fuzzy K-Means is a statistically formalized method [33] which is able to find soft clusters where a particular item can belong to more than one cluster with a certain probability. Like K-Means, this algorithm works with a set of vectors that represent the items and a distance measure to decide to what cluster could belong each item. The basic difference is based on the likelihood estimation for each item, which is used to know the probability to belong to a particular cluster. These membership probabilities are later used by the algorithm to recalculate the new cluster centres.

- **Dirichlet Process Algorithm:** This clustering algorithm performs a Bayesian mixture modelling to generate the resulting clusters [18]. The idea is that a probabilistic mixture of a number of models can be used to explain some observed data, and each observed item is assumed to come from one of these models. Iteratively the items are assigned to the different models using the mixture of probabilities, and the degree of fit, between the item and each model. Once all the items are assigned, new parameters for each model are sampled from the model parameters, considering all of the observed data points that were assigned to the model.

2.3 Complex Networks

Complex Networks have also been used in several domains related to Social Data-Mining [12]. Several of these approaches have been focused on graph models. These networks are usually used to represent a Social Network [23] (i.e. Facebook or Twitter) as a graph, where the users are represented as the nodes of the graph and the relations between them are represented as the edges of the graph. This representation provides a lot of information about the nature of the network (Small-World [30], Random [11], Scale-free [19], etc…), and its main features (strength, paths, authorities, hubs, etc…). It also has been successfully applied to a wide number of different domains such as Marketing [27] and Medicine [21], amongst others.

In this work, it has been applied PageRank [4]. This popular algorithm was defined, and firstly used, to measure the actors influence in any web page. PageRank is a link analysis algorithm initially used by the Google web search engine. It assigns a numerical weight to each element of a linked set of nodes (in the original implementation was used to evaluate the relevance of a particular web page through the number of hyperlinks stored). The main goal is to measure the importance of each node within the graph. The numerical weight assigned to each node n_i is referred to the PageRank value of n_i, and denoted by $PR(n_i)$. The PageRank algorithm is an iterative algorithm which calculates recurrently the following values:

$$PR(n_i) = \frac{1-d}{N} + d \sum_{n_j \in M(n_i)} \frac{PR(n_j)}{L(n_j)} \qquad \textbf{(1)}$$

Where $PR(n_j)$ is the PageRank value of node n_j; d is the damping factor which is used to adjust the algorithm; N is the number of nodes; $L(n_j)$ is the number of out-bound links on node n_j; and $M(n_i)$ is the set of nodes with in-bound links to n_i.

This algorithm is usually solved using an algebraic process or an iterative process. In addition, when the iterative process is used, the PageRank values are usually normalized.

3. METHODOLOGY

To extract collective knowledge from a Social Network, two main different phases have been performed. The first one deals with the *combination of clustering and classification techniques* for sentiment analysis, over a labelled dataset which contains user opinions. In a previous work [3], these data mining techniques were applied separately to extract the trends of user opinions. This work showed that clustering techniques should be helpful for the initial human-labelling process. For this reason, the current 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.

The second one is based on the *analysis of the Social Network structure*, to provide information about how the different communities are constructed, and how the most relevant users of these communities can be found. The following subsections describe these two phases in detail.

3.1 Combination of clustering and classification techniques

Four sequential phases have been performed to combine clustering and classification techniques for the analysis of a text-messages dataset:

1. **Document Preprocessing:** the set of required processes (features extraction, normalization, etc…) in order to feed the clustering and classification algorithms with the target data.
2. **Clustering Process:** several clustering algorithms have been executed using the preprocessed dataset. A comparative study of these algorithms has been carried out using different metrics. And finally, the better algorithm has been selected to extract the main cluster topics.
3. **Re-label the Dataset:** the topics obtained through the clustering process are used to re-label the dataset according to the clusters generated.

4. **Classification Process**: Finally, the classification techniques are applied to compare the results obtained according to the new relabelled dataset with the previous one. A comparative study of the results is carried out to test if the clustering-based trends identification improves the human-labelling work.

3.1.1 Document Preprocessing

Previous Data Mining techniques considered (Classification and Clustering) needs from different kinds of preprocessing. For this reason, two methods have been used according to the nature of those techniques:

- **Data Preprocessing for Classification:** The Preprocessing process consists in some typical steps oriented to simplify 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 for both, to choose the most relevant words and to 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 [11] combined with an Exhaustive Search. The final term-matrix has the following 15 attributes (in Spanish): “bien”, “millones”, “todo”, “#publicidad”, “bonita”, “estas”, “hacer”, “pues”, “quiero”, “toca”, “has”, “llevo”, “mas”, “saben”, “solo”.

- **Data Preprocessing for Clustering:** A very popular model for representing the content of a document or a text is the Vector Space Model (VSM) [13]. Using the VSM, each document is represented by a vector of frequencies of remaining terms within the document. 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. The other function usually used is the Inverse Document Frequency (IDF). Typically, documents are represented as TF-IDF feature vectors. With this data representation, a document represents a data point in a *d*-dimensional space, where *d* is the size of the corpus vocabulary. Text documents are tokenized transforming them into 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 to be used later in the clustering process.

3.1.2 Algorithms Metrics

Any clustering or classification algorithms need a measure to define the distance, or similarity, between the data instances. These measures are defined by the metrics. The metrics which have been considered in this work are the following:

- **Euclidean distance:** Texts or documents are represented as points in a space of several dimensions whose coordinates are based on TF-IDF vector values. Then, it is possible to compute the Euclidean distance [2] between two of these points. For the calculation of this distance between two texts, X and Y, the number of terms will be the number of dimensions N, and the frequency will be coordinates along those dimensions. Therefore, the distance is the square root of the sum of the squares of differences in a position (preference) along each dimension (see Equation 2).

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad \textbf{(2)}$$

  The distance could be computed as 1 / (1 + d_E), so the resulting values are in the range between 0 and 1. When this distance is 0, it means that the texts are identical.

- ***Squared Euclidean Distance Measure:*** The value of this distance measure is the square of the value returned by the Euclidean distance, as Equation 3 shows.

$$d_{E^2}(X,Y) = \sum_{i=1}^{n}(X_i - Y_i)^2 \tag{3}$$

- ***Manhattan Distance Measure.*** In this measure, the distance between any two points is calculated as the sum of the absolute difference of their coordinates [24]. The Manhattan distance between two n-dimensions vectors X and Y is defined as:

$$d_M(X,Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{4}$$

- ***Cosine Distance Measure.*** This measure is based on the uncentred cosine distance [31]. As in the Euclidean distance, texts are represented by points in an n-dimensional space. The distance value will be the cosine of the angle formed between these two term vectors.

$$d_{cos}(X,Y) = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} X_i} \times \sqrt{\sum_{i=1}^{n} Y_i}} \tag{5}$$

When two texts are similar, they have similar term frequencies and therefore they will be close in the space represented. Then the angle formed between these two preference vectors will be very small (near to 0º). In contrast, when the two texts are different, their frequencies vectors will form a large angle. The cosine value is between -1 and 1, where the cosine of a small angle is near 1, and cosine of a large angle (180 degrees) is near -1.

- ***Tanimoto Coefficient Distance.*** This is a distance measure based on the Tanimoto coefficient, or the extended Jaccard coefficient [26]. The definition of the coefficient is the number of common terms sharing by two texts, divided by the number of terms that either texts have in common. Therefore, the coefficient represents the ratio between the size of the intersection and the size of the union of their frequency vectors. The value obtained is between 0 and 1, where if the frequency vectors of the two texts are complete overlapped, the resulting value will be 1.

$$d_T(X,Y) = \frac{\sum_{i=1}^{n}(X_i \wedge Y_i)}{\sum_{i=1}^{n}(X_i \vee Y_i)} \tag{6}$$

- ***Radial Basis Function Kernel (RBF [27]).*** This kernel measure is used to calculate the similarity between two data instances, or points, in the Euclidean Space. The kernel calculates the inverse exponent of the Euclidean Distance. It also uses a control factor (σ) to change the magnitude order. It is defined as:

$$d_{RBF}(X,Y) = e^{\sigma\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}} \tag{7}$$

3.1.3 Evaluation Metrics

Data Mining techniques also require from an evaluation process. These processes are used to validate the models generated by the algorithms. Classification and Clustering techniques usually use different evaluation methods according to their behaviour: clustering is unsupervised, while classification can be evaluated using the data labels in a supervised way.

Due the unsupervised nature of clustering techniques, it is usually extremely hard to carry out reliable and accurate evaluations of the results given from these algorithms. The concept of good partition for a text dataset is sometimes quite subjective. There are two main ways to evaluate the quality of clustering algorithms [34]:

- **Internal:** Two objective functions, intra-cluster and inter-cluster similarity, can be used to evaluate the quality of the generated clusters. The first function, intra-cluster similarity, will try to find low distance values between the documents stored in the cluster. This means that those documents grouped in the same cluster are similar. The second function, inter-cluster similarity, will look for high distances values in the documents that belong to different clusters. This means that documents grouped in different clusters are really dissimilar.

- **External:** Compare the clustering results with a trusted manual categorization.

Due to the fact that the dataset used in this analysis has been labelled by human experts, these two evaluation criteria (internal and external) will be considered in the experimental analysis.

Classification and the external clustering evaluations have been carried out using the classical metrics of Precision, Recall and F-measure [13], which are defined as follows:

- **Precision**: In the area of Information Retrieval, *precision* is used to represent the fraction of retrieved documents that are relevant to the search. In our approach, *precision* will be used to represent how many instances have been correctly classified in a cluster (including correct instances, or *true positives*, and incorrect classified instances, or *false positives*).

$$Precision = \frac{tp}{tp + fp} \tag{8}$$

- **Recall**: In Information Retrieval, the *recall* value represents the fraction of the documents that are relevant to the query that are successfully retrieved. In our approach, the *recall* value is used to measure how many instances have been correctly grouped in the same cluster, from the whole number of instances that should belong to this class.

$$Recall = \frac{tp}{tp + fn} \tag{9}$$

- **F-measure**: This metric is the *harmonic mean* between Precision and Recall values.

$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

3.2 Analysis of Social Network structure

The Social Network analysis is based on the following steps:

- First, **the Network is represented**. In this case, the users have been considered as the network nodes and their relationships are the edges. The relationships which have been considered in this work are: Re-Tweets and mentions, i.e., when a user re-tweets a message from the other user, or mentions the other user in any of its tweets, the edge is generated. The arrow corresponds to the user who is re-tweeting or mentioning to the other user. This decision has been taken because the users are usually divided into two kinds of users according to the literature: Authorities (those users who are followed by several users) and Hubs (those users who follow several users).

- Second, the information about **the user opinions is represented** in the network. This information is taken from the tweet labels which have been generated after the clustering analysis. Using this information, we are able to distinguish what kind of opinion is more propagated in the network and how the communities are generated.

- Finally, the PageRank algorithm is applied to the communities to **discriminate the most relevant users** or actors within each community. This information determines the most relevant user of the community which is important for marketing studies.

4. EXPERIMENTAL RESULTS

This section describes the dataset used to validate the methodology proposed in this work, the experimental setup made in the Data Mining algorithms, and provides a detailed discussion on the results obtained.

4.1 Dataset Description

The data which have been analysed in this work come from Twitter. Twitter is a Social Network where people usually publish information about personal opinions. It is divided into two kinds of user behaviours: follower and following. As a follower, the user receives information of people who are followed by him, and as a following, the user information is sent to his 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 this analysis is based on 100 comments about IKEA®. The comments have been extracted from “02-11-2013 15:24” to “02-18-2013 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:

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

4.2 Parameter Selection

In any Data Mining (DM) process, the parameter selection is one of the first (and critical) steps that must be done before any DM algorithm or method can be applied over the data. The identification, and selection, of the relevant parameters allows guiding correctly the algorithm, avoids misclassifications, and provides reliable and accurate results.

4.2.1 Clustering setup

A clustering process is related to the problem of organizing items from a given collection into groups with similar characteristics called clusters. This process involves two main features, the clustering algorithm applied (the method used to group the similar items together), and the distance metric used to measure the similarity among these items. It is difficult to decide which clustering configuration is the best (how many clusters to generate or what kind of similarity measure to choose). Therefore, an evaluation and study of the clusters quality is required. Firstly, the three clustering algorithms explained in Section 2 (K-Means, Fuzzy K-Means and Dirichlet) have been executed for each distance detailed in Section 3 (Cosine, Tanimoto, Manhattan and Euclidean) using the dataset described in the previous section.

However, as it was described in Section 2, some of those algorithms (K-means) need to fix parameters, such as the number of clusters (K) to be identified in the data. From our initial previous results [3] using Tweets, where we tried to classify the content of these messages, K value was fixed to 5. Using this value, all of the algorithms considered (with the different distances measures) were executed over the data to look for the best and most accurate clustering algorithm. These results are shown in Table 1.

Table 1**.** Results from the clustering algorithms considered. These algorithms have been applied to the topic problem detection using four distance measures.

Algorithm	Distance Measure	Intra Cluster Dis.	Inter Cluster Dis.
K-means	Cosine	0.54	0.48
	Tanimoto	**0.58**	**0.53**
	Manhattan	0.63	0.49
	Euclidean	0.60	0.42
Fuzzy K-means	Cosine	0.63	0.56
	Tanimoto	0.59	0.53
	Manhattan	0.59	NaN
	Euclidean	0.62	0.31
Dirichlet	Cosine	0.63	0.45
	Tanimoto	0.61	0.39
	Manhattan	0.61	0.50
	Euclidean	0.60	0.45

Analysing the clustering results shown in Table 1, it can be noticed that K-means algorithm obtains better results. The algorithm's goal is to create clusters that are coherent internally, but clearly different from each other. This means that documents within a cluster should be as similar as possible (*lower* intra cluster distance), and documents in one cluster should be as dissimilar as possible from documents in other clusters (*higher* inter cluster distance). Regarding the intra and inter cluster metrics (for K-Means), the distance measure with a low intra cluster value and the higher inter cluster value is the *Tanimoto* distance.

Other algorithms, such as Fuzzy K-Means, obtain good intra-cluster distances but with poor inter-cluster distances results. This may be due to the nature of the algorithm, Fuzzy K-Means, trying to generate clusters that could be overlapped, so one item (or Tweet) could belong to several classes. However, our current dataset needs from a partitional classification, given from the marketing experts, in four non-overlapped categories.

Once the best algorithm and distance measure have been selected (*K-Means* using *Tanimoto* distance), it is necessary to evaluate the best number of clusters (*K*) for the new dataset. To achieve this goal, the algorithm was executed with a value of *K* from 2 to 10, and the best results obtained are shown in Figure 1.

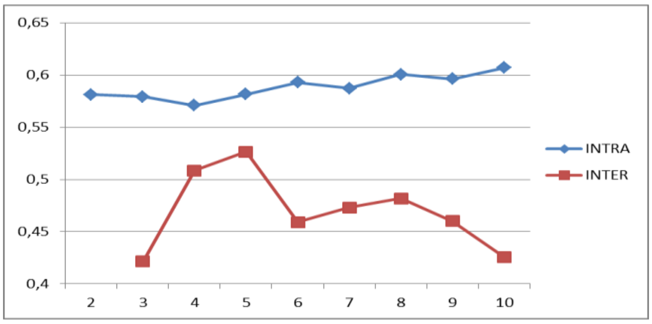


Fig. 1. Distance comparison using K-Means with Tanimoto distance for K between 2 and 10.

As Fig. 1 shows, the best results were achieved with K equals to **5.** The inter-cluster distance is maximized and the intra-cluster measure takes a low value. The inter-cluster distance improves their results until K equals to 5 is reached, and for higher values of K, the results dramatically become worse. Therefore, the clusters found in this solution can be better differentiated, and they have been chosen to extract the message topics and to make a more detailed analysis of the user opinions.

Finally, the other interesting conclusion from these experiments is related to the best obtained value from the clustering algorithm. Although the human experts classified in four different categories analysed Tweets, our best results show five (K=5) different clusters, or categories, although K=4 provides goods results too, with a better intra-cluster distance (closer to the value of K=5), the intra-cluster distance for K=4 was clearly worse. For this reason, these five groups, or categories, will be considered in the next step of our methodology.

4.2.2 Classification setup

Once the clustering results have been obtained, the classification algorithms are applied to the new class discrimination to compare the old classification results with the new results generated by this methodology. The algorithms which have been used are Naïve Bayes, C4.5, SVM and K-Nearest Neighbour, as mentioned in Section 2. As with clustering algorithms, these new algorithms need several parameter selections. Table 2 shows this parameter selection. The execution and results from the execution of classification algorithms are shown in Section 4.3.

Table 2. Parameter selection for the classification algorithms.

Algorithm	Parameters	Metric
Naive Bayes	-	-
C4.5	Confidence factor = 0.25	Information Entropy [22]
	Min. Number objects = 2	
SVM	$\sigma = 0.1$	RBF
K-Nearest Neighbour	K = 5	Euclidean Distance

4.3. Clustering Results

The clustering algorithm with the set up selected (5-Means using Tanimoto distance) is applied on the dataset to extract the conversation topics based on clusters. Table 3 shows the result clusters and the topics extracted from them. Also in Fig. 2, the clusters obtained are plotted in a graph representation to provide a better appreciation of the cluster structure and size.

Table 3. Topics extracted for the clustering process (5-Means using Tanimoto distance)

Cluster Num	Class	Topics	Colour
1	Dissatisfaction	t.co, http, mueble, meses, dentro	🟥
2	Dissatisfaction	ir, comprar, quiero, saber, mas	🟦
3	Neutral	casa, muebles, dos, cada, http	🟩
4	Satisfaction	familia, piso, tener, nuevo, ganas	🟨
5	Satisfaction	comprando, voy, horas, tarde, ponemos	🟪

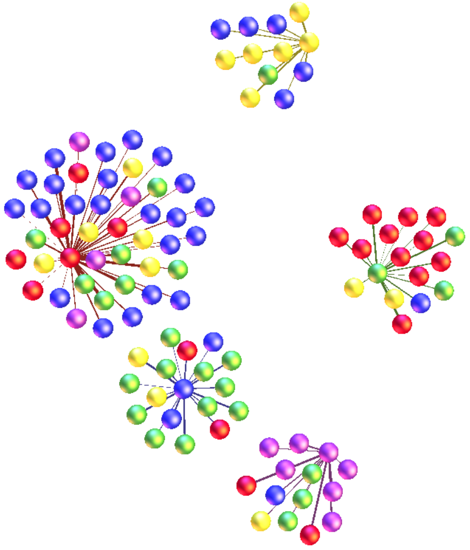


Fig. 2. Graph representation of cluster results for 5-means using Tanomino distance.

Analysing the number of clusters related to each class, there are various remarkable aspects. The Dissatisfaction and Satisfaction classes are separated in two sub-trends per class ("Dissatisfaction1"-"Dissatisfaction2" and "Satisfaction1"-"Satisfaction2" respectively). It means that a more detailed analysis of these trends would perform a better separation of the user opinions.

The Exclusion class is undistinguishable in all cases. It means that this class should not be considered as a trend in the Tweets ("Neutral" label). Therefore, according with the topics obtained through the clustering process, the dataset is relabelled (see Table 4).

Table 4. Re-labelled dataset using the clustering topics obtained.

Old Class	Matching Cluster	New Class
Exclusion	4	Satisfaction1
Neutral	3	Neutral
Satisfaction	4, 5	Satifaction1 Satisfaction2
Dissatisfaction	1, 2	Dissatisfaction1 Dissatisfaction2

Table 4 shows how the "Old Class" categories, generated from marketing experts, the new labels generated ("New Class"), and finally the correspondence between these new labels and the clusters that were identified from the clustering algorithm ("Matching Cluster").

4.4. Classification Results

Once the clustering results have been used to generate the new classes, the classifiers have been applied to the new class selection to compare the results. Tables 5 and 6 show the classification results for the new and old classes' discrimination, respectively.

Table 6 shows that the best accuracy results for the last classes selection was obtained by the Naïve Bayes algorithm with an average F-measure of **0.549**. However, this value has been improved with the new class distribution. Using the proposed methodology, NB obtains an average F-measure value of **0.567** for the 5 classes defined by the clustering analysis.

Table 5. Results for the classification methods applied to the **new classes** generated by the clustering analysis.

Technique	Class	Precision	Recall	F-measure
NB	Satisfaction1	0.667	0.435	0.526
	Satisfaction2	0.458	0.688	0.55
	Neutral	1	1	1
	Dissatisfaction1	0.38	1	0.551
	Distatisfaction2	0.609	1	0.757
	Average			0.567
KNN	Satisfaction1	0.526	0.435	0.476
	Satisfaction2	0.808	0.553	0.656
	Neutral	0.429	0.214	0.286
	Dissatisfaction1	0.4	0.25	0.308
	Distatisfaction2	0.211	0.889	0.34
	Average			0.4132
C4.5	Satisfaction1	0.875	0.609	**0.718**
	Satisfaction2	0.609	0.875	**0.718**
	Neutral	1	1	**1**
	Dissatisfaction1	1	0.222	**0.364**
	Distatisfaction2	0.667	1	**0.8**
	Average			**0.72**
SVM	Satisfaction1	0.44	0.478	0.458
	Satisfaction2	0.667	0.5	0.571
	Neutral	0.775	0.816	0.795
	Dissatisfaction1	0.143	0.111	0.125
	Distatisfaction2	0.438	0.5	0.467
	Average			0.483

Table 6. Results for the classification methods applied to the **old classes**.

Technique	Class	Precision	Recall	F-measure
NB	Neutral	0.61	0.973	**0.75**
	Satisfaction	0.652	0.484	**0.556**
	Dissatisfaction	0.692	0.391	**0.556**
	Exclusion	0.5	0.25	**0.333**
	Average			**0.549**
KNN	Neutral	0.605	0.703	0.65
	Satisfaction	0.5	0.387	0.436
	Dissatisfaction	0.5	0.478	0.489
	Exclusion	0.2	0.25	0.222
	Average			0.449
C4.5	Neutral	0.45	0.973	0.615
	Satisfaction	0.714	0.323	0.444
	Dissatisfaction	1	0.174	0.296
	Exclusion	0	0	0
	Average			0.339
SVM	Neutral	0.621	0.973	0.758

Satisfaction	0.571	0.516	0.542
Dissatisfaction	0.769	0.435	0.556
Exclusion	0	0	0
Average			0.464

The best classifier of the new classes' discrimination is C4.5 which obtains an average F-measure value of **0.72**. This value is an important improvement compared with the previous results. The class "Neutral" is clearly discriminated in this new analysis; also, all the classes have a good discrimination for all the algorithms. Although there are more classes in the new analysis, the algorithms are able to discriminate them easier, so this provides an accurate result for the final evaluation.

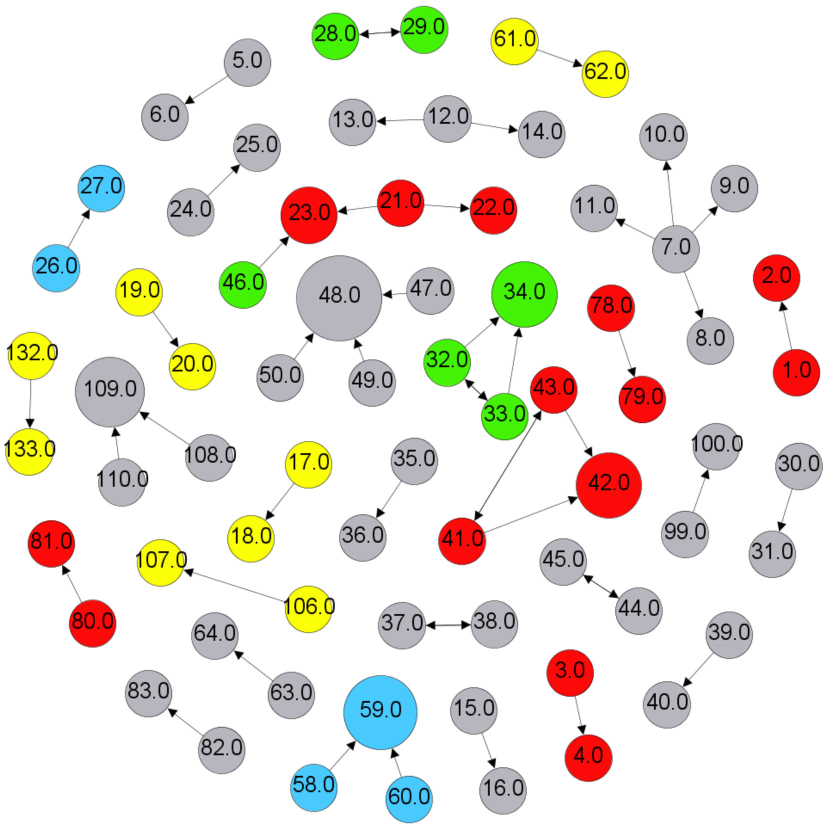
All the classifiers obtain better results with the classes than with the old ones, except KNN which obtains worse results than in the previous discrimination. It should be because the number of classes is higher and this algorithm is based on the closest neighbours.

4.5 Social Network Results

The Social Network analysis of the Tweets dataset has been applied to 100 Tweets and all their re-tweets extracted during the dates of the extraction process. The re-tweets set contains 36 extra instances. The Network has been created using the users as nodes, and the re-tweets as their relationships. The analysis of the network is focused on the identification of influential actors inside the communities generated by the users.

A relevant subset of the communities formed by the Social Network generated is represented in Fig. 3. This figure shows a plot of the Social Network generated by the re-tweets and users mentions. There are some details that need to be explained before proceeding with the Network analysis:

- This Network has been generated using the re-tweets and user mentions. For example, when user 1 mentions or re-tweets user 2, an arrow from user 1 to user 2 is added to the network.
- The arrows represent those users which have been re-tweeted or mentioned, for example, user *"48.0"* (grey big node, near to the centre of the figure) has been re-tweeted or mentioned three times.
- The node size corresponds with the PageRank application results. Those nodes which are bigger have higher PageRank value.



Nodes			
🟥	Satisfaction1	🟩	Satisfaction2
🟦	Dissatisfaction1	🟨	Dissatisfaction2
⬜	Neutral		

Fig. 3. Representation of the Social Network generated by the re-tweets and mentions of the users. The colours represent the user opinions and the node size represents the PageRank value for each node.

The analysis shows that there are several communities which connect two single nodes. The biggest communities formed in the communication are related to Satisfaction and Neutral opinions. Dissatisfaction has several communities with only two nodes. The dissatisfaction opinions are not well propagated in the network; however neutral and satisfaction opinions are more propagated. The community size is low because the number of tweets is small (100 of tweets) and the number of users (including mentions) is high (133 users).

The most interesting communities are those formed by users:

- **46.0, 23.0, 21.0, 22.0**: This community has an important actor (node 23.0) and the information about the brand which is propagated is satisfactory.
- **50.0, 48.0, 47.0, 49.0**: This community is focused in one main actor (node 48.0) which has generated an important propagation centre.
- **32.0, 33.0, 34.0**: This community forms a triangle between the nodes. The most relevant actor is 34.0 and also propagates satisfactory opinions.
- **41.0, 42.0, 43.0**: This community is similar to the last and also propagates satisfactory opinions.
- **7.0, 8.0, 9.0, 10.0, 11.0**: This community is formed by a hub (node 7.0) which takes information of several authors.

This information is helpful to conclude that the most important opinions about IKEA® are neutral and satisfactory opinions.

5. CONCLUSION

This work has shown the application of Data Mining and Complex Network 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. On the one hand, clustering and classification techniques have been combined to extract the trends of user opinions and also improve the classification results through an initial clustering analysis. On the other hand, Complex Network analysis has been used to study the communities formed by the users and their interactions.

Clustering techniques are applied to extract the main topics and generate the topic clusters. Using the information obtained from these clustering techniques, the Tweet dataset which has been classified is relabelled according to the clusters generated by the clustering process. Finally, a comparative classification study with our previous work has been presented. The combination of clustering and classification techniques has achieved better results than the use of simple classification algorithms. Hence, this new methodology shows that clustering techniques provide more detailed information about the collective trends, and these techniques are helpful to guide the initial human-labelling process.

The relabelled dataset is also used for the Complex Network analysis. These techniques are used to extract the communities formed by the users, studying the opinion propagation through the communities and also measuring the importance of the users inside the Social Network. This work shows the practical utility on the combination of those Data Mining techniques with Complex Network methods, to automatically discover knowledge and collective trends in textual data extracted from Twitter.

ACKNOWLEDGMENTS

The preparation of this manuscript has been supported by the Spanish Ministry of Science and Innovation under the following projects: TIN2010-19872 and ECO2011-30105 (National Plan for Research, Development and Innovation), as well as the Multidisciplinary Project of Universidad Autónoma de Madrid (CEMU2012-034). The authors thank Ana M. Díaz-Martín and Mercedes Rozano for the manual classification of the Tweets.

REFERENCES

[1] H. Ahonen-Myka. “MIRining 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*, New York, NY, USA. ACM, 2005 pp. 255–256.

[2] D. G. Bailey. “An efficient euclidean distance transform”, in *Proceedings of the 10th InternationalWorkshop in Combinatorial Image Analysi, IWCIA 2004*, Auckland, New Zealand, Springer, 2004, pp. 394–408.

[3] G. Bello-Orgaz, H. Menéndez, S. Okazaki, and D. Camacho. “Extracting Collective Trends from Twitter using Social-based Data Mining”, in *Proceedings of the 5th International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI 2013*, Craiova, Romania. Springer-Verlag 11-13 September 2013.

[4] S. Brin and L. Page. “The anatomy of a large-scale hypertextual web search engine”, in *Proceedings of the seventh international conference on World Wide Web 7, WWW7*, Amsterdam, The Netherlands. Elsevier Science Publishers B. V., 1998. pp. 107–117.

[5] A. Qazi, R. G. Raj, M. Tahir, M. Waheed, S. U. R. Khan, and A. Abraham, “A Preliminary Investigation of User Perception and Behavioral Intention for Different Review Types: Customers and Designers Perspective”, *The Scientific World Journal*, Vol. 2014, Article ID 872929, 8 pages, 2014. doi:10.1155/2014/872929.

[6] C. Cortes and V. Vapnik. “Support-vector networks”. *Machine Learning*, Vol. 20, 1995, pp. 273–297.

[7] L. Y. Wei, R. Mahmud and R. G. Raj, “An application of case-based reasoning with machine learning for forensic autopsy”, *Expert Systems with Applications*, Vol 41, No. 7, 2014, pp. 3497-3505, ISSN 0957- 4174, http://dx.doi.org/10.1016/j.eswa.2013.10.054. (http://www.sciencedirect.com/science/article/pii/S0957 417413008713).

[8] 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*. New York, NY, USA. ACM, 1992. pp. 318–329,

[9] P. Domingos and M. Pazzani. “On the optimality of the simple bayesian classifier under zero-one loss”. *Machine Learning*, Vol. 29, No 2-3, Nov. 1997, pp. 103–130.

[10] P. Domingos and M. Richardson. “Mining the network value of customers”, in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01)*. New York, NY, USA, ACM, 2001, pp. 57-66.

[11] P. Erdös and A. Rényi. “On random graphs”. *I. Publ. Math. Debrecen*, Vol 6, 1959, pp. 290–297.

[12] V. Balakrishnan, F. G. Sim and R. G. Raj. “A one-mode-for-all predictor for text messaging”, *Maejo International Journal of Science and Technology*, Vol. 5, No. 2, 2011, pp. 266-278.

[13] W. B. Frakes and R. A. Baeza-Yates. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall, 1992.

[14] J. J. Jung, J. Euzenat. “Towards semantic social networks”, in *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*. Berlin, Springer Berlin Heidelberg, 2007, pp. 267-280.

[15] J. J. Jung, J.-S. Yoon, G.-S. Jo. “Collaborative information filtering by using categorized bookmarks on the web”, in *Proceedings of the 14th international conference on Web knowledge management and decision support*. Springer Berlin Heidelberg, Vol 2543, 2003, pp. 237-250.

[16] J. A. Hartigan and M. A. Wong. “A K-means clustering algorithm”. *Applied Statistics*, Vol. 28, 1979, pp. 100–108.

[17] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, Mar. 1990.

[18] D. T. Larose. *Discovering Knowledge in Data*. John Wiley and Sons, 2005.

[19] Albert-László Barabási and Réka Albert. “Emergence of Scaling in Random Networks”. *Science*, Vol. 286, No. 5439, October 1999, pp. 509–512.

[20] Y. Li, S. M. Chung, and J. D. Holt. “Text document clustering based on frequent word meaning sequences”. *Data Knowl. Eng.*, Vol. 64, No. 1, Jan. 2008, pp. 381–404.

[21] L. Thompson, K. Dawson, R. Ferdig, E. W. Black, J. Boyer, J. Coutts, and N. P. P. Black. “The intersection of online social networking with medical professionalism”. *Journal of general internal medicine*, Vol. 23, No. 7, July 2008, pp. 954–957.

[22] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University

Press, New York, NY, USA, 2008.

[23] M. E. J. Newman. “The Structure and Function of Complex Networks”. *SIAM Review*, Vol. 45, No. 2, 2003, pp. 167–256.

[24] T. Nis. *Dictionary of Algorithms and Data Structures*. U.S. National Institute of Standards and Technology, 2005.

[25]. J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[26] V. Sachdeva, D. M. Freimuth, and C. Mueller. “Evaluating the jaccard-tanimoto index on multi-core architectures”, in *Proceedings of the 9th International Conference of Computational Science, ICCS 2009,* Baton Rouge, LA, USA. Springer, May 25-27, 2009, pp. 944-953.

[27] T. Smith. “The social media revolution”. *International Journal of Market Research*, Vol. 51, No. 4, July 2009, pp. 559–561.

[28] J. J. Jung. “Contextual Synchronization for Efficient Social Collaborations in Enterprise Computing: a Case Study on TweetPulse”, *Concurrent Engineering – Research and Applications*, Vol. 21, No. 3, 2013, pp. 209–216.

[29] J. J. Jung. “Cross-lingual Query Expansion in Multilingual Folksonomies: a Case Study on Flickr”, *Knowledge-Based Systems*, Vol. 42, 2013, pp. 60–67s.

[30] D. J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton Studies in Complexity, 1999.

[31] H. Xiong. “Hyperclique pattern discovery”. *Data Mining and Knowledge Discovery Journal*, Vol. 13, No. 2, September 2006, pp. 219-242.

[32] 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*, New York, NY, USA. ACM, 1998, pp. 46–54.

[33] D. Zhang and S. Chen. “Fuzzy clustering using kernel method”, in *Proceedings of the International Conference on Control and Automation, ICCA*, Xiamen, China. 2002, pp. 123–127.

[34] Y. Zhao and G. Karypis. “Empirical and theoretical comparisons of selected criterion functions for document clustering”. *Machine Learning Journal*, Vol. 55, No. 3, June 2004, pp. 311–331.