Automatic Categorisation of Comments in Social News Websites

Igor Santos*, Jorge de-la-Peña-Sordo, Iker Pastor-López, Patxi Galán-García, Pablo G. Bringas

Laboratory for Smartness, Semantics and Security (S^3Lab), University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain
Telephone: +34944139003 Fax: +34944139166

Abstract

The use of the social web has brought a series of changes in the way how content is created. In particular, social news sites link stories and the different users can comment them. In this paper, we propose a new method based on different features extracted from the text able to categorise the comments. To this end, we use a combination of syntactic, semantic and opinion features and machine-learning classifiers to classify a comment within 3 different categorisation types: the focus of the comment, the type of information contained in the comment and the controversy level of the comment. We validate our approach with data from 'Menéame', a popular Spanish social news site.

Keywords: spam detection, information filtering, content filtering, machine-learning, web categorisation

1. Introduction

The Web has evolved over the years and, now, not only the administrators of a site generate content. Users of a website can express themselves and make content available in sites that show their feelings or opinions about a

*Corresponding author

Email addresses: isantos@deusto.es (Igor Santos), jorge.delapenya@deusto.es (Jorge de-la-Peña-Sordo), iker.pastor@deusto.es (Iker Pastor-López), patxigg@deusto.es (Patxi Galán-García), pablo.garcia.bringas@deusto.es (Pablo G. Bringas)

fact. Therefore, users can now rapidly publish content and this content is emerging in the Web.

Social news websites such as Digg[1] or 'Menéame'[2] are popular social websites. These sites work in a very simple and intuitive way: users submit links to stories online, and other users of those systems rate them by voting their news. Most voted stories appear, finally, in the frontpage (Lerman, 2007).

In this work, we focus on 'Menéame'. This social news website has already a method for automatic moderation of comments and stories in order to filter them. However, it is based on the votes of other users and, therefore, it may not be objective. In a similar vein, there are approaches to filter spam in reviews (Jindal & Liu, 2007, 2008). The authors proposed a method based on several opinion and syntactic features to automatically filter spam messages in product reviews in the website 'Amazon'[3].

Given this background, we propose the first approach that is able to automatically categorise comments in these social news sites.

This approach could be used in any type of web content that allows users to comment or refer to other content in the Internet. It can be also used in order to modify the content of a page in order to make it suitable for different kinds of users, filter inappropriate content or to categorise users with regards to the content they generate.

The approach employs different syntactic, semantic, statistical and opinion features to build a representation of the comments. Based on this representation, machine-learning-based classifiers are trained to categorise the comments. To this end, we concentrate on three possible types of classifications: the focus of the comment (i.e., if the comment focuses on the news story or on another comment), the type of information (contribution, irrelevant or opinion) and the controversy level of the comment (normal, controversial, very controversial or joke).

Summarising, our main contributions are:

- A new method for representing comments in social news websites.
- A machine-learning-based method for categorising comments in social news sites.

[1] `http://digg.com/`
[2] `http://meneame.net/`
[3] `http://www.amazon.com/`

- We show that these methods can achieve high accuracy rates in three different classification tasks with data extracted from 'Menéame'.

The remainder of this paper is organised as follows. Section 2 describes in detail our proposed method. Section 3 describes the experiments performed and presents results. Section 4 discusses the main limitations of this work and outlines the avenues of the future work.

2. Method Description

2.1. Data from Meneame.net

'Menéame' is a Spanish social news website, in which news and stories are promoted. It was developed in later 2005 by Ricardo Galli and Benjamín Villoslada and it is currently licensed as free software. At the beginning, it was focused on scientific and technological topics, but nowadays it is open to any topic such as politics, society or sports. Also, as the number of the users of 'Menéame' grew, so did the quality and quantity of the contributions.

Any user (even if it is not registered in the system) can vote the news stories in the front page or in the pending section, which are news that have not been contrasted yet. Registered users can send news to the system. A news story is held in the pending queue. There, the story will be voted by different readers or users. Registered users can also make a negative vote and comment the news story.

'Menéame' ranks their users depending on their 'karma'. The 'karma' is a value between 0 and 20. When a new user is registered a value of 6 point of 'karma' is given. 'Karma' is computed based on the performed activity in the previous 2 days. To this end, the algorithm combines 4 different components: positives votes received of the sent news, positive votes made, negative votes made and votes received of a user's comments. When a news story is in the pending queue, the 'karma' of the users that vote the story are added to its value and if they surpass a threshold they are published in front page. Otherwise, the stories that accumulate negative votes, will be sent to the discarded section. Usually, these contributions are either irrelevant, old, bothering, sensationalist, spam, replicated, micro-blogging, mistaken or plagiarism.

The possible ranks that 'Menéame' gives to their users are:

- **Normal:** The normal user is every user that is registered in the site and starts with a 'karma' of 6.

- **Special:** When a normal user's 'karma' surpasses the 80% of the maximum value of 'karma', the user becomes special. These users can edit news which are in the pending queue. They can lose until the 60% of their 'karma', then, they come back to be normal.

- **Blogger:** This category is reserved to users that have made significant contributions. Their privileges are the same that the special users have, but they can also discard news. This status is never lost.

- **Admin:** These users' task is digital promotion. They have the same privileges as the bloggers.

- **God:** A god user has the same privileges that the admin users and they can also view other users' profiles. They are also the only type of users that can edit comments.

None of the users is able to edit or remove the 'karma' of the stories neither edit their number of votes.

Besides, there are 2 possible special status: disabled and auto-disabled. If a user abuses of the system, the user will be ranked with the disabled status. When a user by him/herself wants to stop using the system, the user's status will be auto-disabled.

The sending phase has no moderation, but some guidelines are given as advice in order to avoid negative votes. For instance, avoid using caps or exclamation marks, make the titles match, put the story in its proper category, provide the link to the original article and so on. 'Menéame' express in their terms of use how news should be submitted[4]: *"The title, snippet, geolocation, and tags, as well as the category in which the news story is inserted, must reflect and should not distort the content of the linked newsstory. 'Menéame' is not a microblogging site and it is not intended to generate news or opinions in the description of the story."*

Figure 1 shows the structure of a story when it is in the front page. The title of the news story should be the same that the one in the external story. After the title, the user links the story. A description of the story has to be written that should be descriptive about the story. In the bottom of the news story, we can notice the number of comments, the value of the 'karma'

[4]Extracted from `http://www.meneame.net/legal.php`

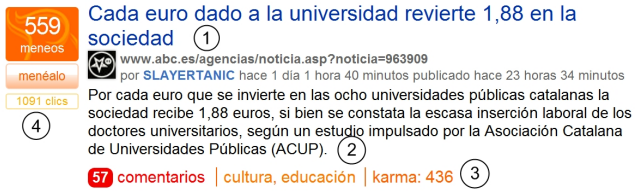


Figure 1: Structure of a story in 'Menéame': (1) is the title of the story; (2) is the description of the new; (3) indicates the number of comments, the 'karma' and the tags; and (4) indicates the number of votes.

and the tags. Besides, in the left side, the number of positive votes of the news story is displayed.

Figure 2: Structure of a story in 'Menéame' when comments are displayed: (1) refers to the tags of the story and (2) details the votes of the news story.

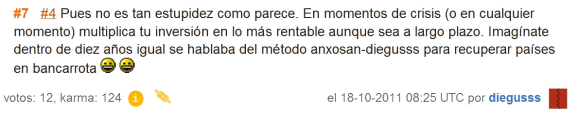
Figure 2 shows the structure of a story the comments are displayed. In addition to the data in the front page, the tags of the story are displayed as well as the votes are detailed in their different categories.

Figure 3 shows a comment in 'Menéame'. The first thing that appears is the number of the comment, which in this case is seven. Next, another number appears that references another comment. In this case the user is giving an opinion about a previous comment.

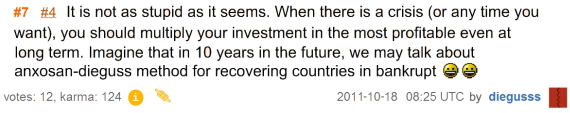
We categorise the comments in three different classifications. In order to make the explanation clearer, we show actual examples from of the different categories for each of the different classifications. These examples have been taken from the story shown in Figure 4.

Each one of the three different classifications have several possible classes. They are the following ones:

- **Type of Information:** The type of information indicates what the user is doing in its comment. It can be:



(a) Example of a comment in Spanish.



(b) Translated comment.

Figure 3: Example of a comment.

Title: The Government has been asked to prevent an 'atheist procession' the Maundy Thursday which will be at the same time that the traditional one in Lavapiés

Description: 'HazteOir' group, the political party 'Alternativa Española', churches like 'Santo Miguel Arcángel' and other catholic collectives asked the Government in Madrid to not allow the 'atheist procession' on Maundy Thursday.

Tags: Church, imposition, atheism, prohibition, government

Figure 4: An example of a news story.

  - *Contribution:* The user contributes by adding new information. Figure 5 shows an example of a contribution comment in the previous story.
  - *Irrelevant:* These comments do not contribute to the main article neither to others previous comments. Figure 6 shows an irrelevant comment.
  - *Opinion:* These comments express the user's particular opinion about the topic discussed in the story. Figure 7 shows an opinion.

- **Focus of the comment:** The comment can be focused either on the main story or on another comment. Figure 8 shows an example of a comment that focuses on the main story whilst Figure 9 shows an

#201 #191 By the way, I forgot telling you that in that town I told you before, the one where a church installed a megaphone to say the prayers before the Sunday service and made every inhabitant to get up early even if they don't want to, the major was of the political party PSOE and he didn't do anything to solve it. There is a clear contradiction between the public agenda of his political party and his actions, but some people do like their chair, don't they?

votes: 0, karma: 8, 18 hours and 23 minutes ago by strel

Figure 5: An example of a contribution.

#225 Good luck, friends!

votes: 0, karma: 6, 12 hours and 46 minutes ago by pavlenka

Figure 6: An example of an irrelevant comment.

example of a comment focused on another comment. Although the comments that refer to another comment contain the 'comment ID' of the cited one, it might happen that the comments remit another comment to support or discuss the news. Therefore, a simple parser searching the character # may not be as appropriate as the extracted features described in sub-section 2.2.

- **Controversy Level:** We categorise the controversy level in three de-

#205 #70 I agree with you, and moreover, I think they should do it even harder, because so much it has cost us to be free of a belief or, at least, reduce it to, now, let them play with our morality and, also, for sociocultural reasons, to be expanded in the future (there is a rapid growth of the population professing such belief)

votes: 1, karma: -1, 18 hours 1 minute ago by RaistlinMajere

Figure 7: An example of an opinion.

#202 The same that always happens: one gets disturbed because he/she wants, no because he/she should. If they concentrate more, they would realise that there are different stuffs than themselves and their truth but that damn egocentric point of view, thinking that everyone is against me...

votes: 0, karma: 6, 18 hours 13 minutes ago by Natxo-Pistatxo

Figure 8: An example of a comment focusing in the main story.

#204 #201 Then, you cannot be affiliated to PSOE and be a religious guy?... Indeed, taking into consideration that the PSOE is a center-right party, I do not really get where is the surprising fact. Would you understand it if it was from the PP?

votes: 0, karma: 6, 18 hours 8 minutes ago by xaphoo

Figure 9: An example of a comment focusing in another comment.

grees: normal, controversial and highly controversial and, also, an additional one that is used for funny or ironic comments.

- *Normal:* A normal comment is the one that raises no controversy or irony.
- *Controversial:* A comment that, on purpose, seeks controversy with an harmful tone. Figure 10 shows an example of a controversial comment.

#218 A lot of people being proud of their atheism, and the 90% of them made gifts to their children in Xmas..., and rate this comment negative if you dare, I don't mind.

votes: 0, karma: 6, 14 hours 46 minutos ago by canaam

Figure 10: An example of a controversial comment.

– *Very Controversial:* It seeks to create controversy in an exaggerated way, being hurtful or disrespectful. In other words, a troll user. Figure 11 shows a very controversial comment.

> #206 #10 You say that there have not been protests against the use of the burka in Spain? And not only by atheists, of course. When we have the same number of asshole fundamentalist both Christians and Muslims in Spain, we will be the first ones to proclaim our stupidity to the four winds. We are not afraid either of you or them.
>
> votes: 0, karma: 6, 17 hours 57 minutes ago by Despero

Figure 11: An example of a very controversial comment.

– *Joke:* These comments try to make a joke and be funny. Figure 12 shows an example of a funny comment.

> If my colleagues and me walk through the streets handing out oil, grease, wax or other products likely to cause slips, we will have problems. The least they will call us will be vandals or we will be fined, or both.
>
> votes: 0, karma: 6, 13 hours 54 minutes ago by pacorron

Figure 12: An example of a joke comment.

2.2. Extracted Features

In this sub-section, we detail which features we extract from the comments. We divide these features into 3 different categories: opinion, statistical and syntactic features.

2.2.1. Statistical Features

The statistical feature category has several features we have used:

- **Comment body:** We used the information contained in the body of the comment. To represent the comments we have used an IR model.

An IR model can be defined as a 4-tuple $[\mathcal{C}, \mathcal{Q}, F, R(q_i, c_j)]$ (Baeza-Yates & Ribeiro-Neto, 1999) where $\mathcal{C}$, is a set of representations of comments; F, is a framework for modelling comments, queries and their relationships; $\mathcal{Q}$, is a set of representations of user queries; and, finally, $R(q_i, c_j)$ is a ranking function that associates a real number with a query q_i ($q_i \in \mathcal{Q}$) and a comment representation c_j, so that ($c_j \in \mathcal{C}$).

As $\mathcal{C}$ is the set of comments c, $\{c : \{t_1, t_2, ...t_n\}\}$, each comprising n terms $t_1, t_2, \ldots, t_n$, we define the weight $w_{i,j}$ as the number of times the term t_i appears in the comment c_j if $w_{i,j}$ is not present in c, $w_{i,j} = 0$. Therefore, a comment c_j can be represented as the vector of weights $\vec{c_j} = (w_{1,j}, w_{2,j}, ...w_{n,j})$.

On the basis of this formalisation, IR systems commonly use the Vector Space Model (VSM) (Baeza-Yates & Ribeiro-Neto, 1999), which represents documents algebraically as vectors in a multidimensional space. This space consists only of positive axis intercepts. Documents are represented by a term-by-document matrix, where the $(i, j)^{th}$ element illustrates the association between the $(i, j)^{th}$ term and the j^{th} comment. This association reflects the occurrence of the i^{th} term in comment j. Terms can represent different textual units (e.g., words or phrases) and can also be individually weighted, allowing the terms to become more or less important within a given comment or the comment collection $\mathcal{C}$ as a whole.

We used the *Term Frequency – Inverse Document Frequency* (TF–IDF) (Salton & McGill, 1983) weighting schema, where the weight of the i^{th} term in the j^{th} document, denoted by $weight(i, j)$, is defined by:

$$weight(i, j) = tf_{i,j} \cdot idf_i \tag{1}$$

where *term frequency* $tf_{i,j}$ is defined as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

where $n_{i,j}$ is the number of times the term $t_{i,j}$ appears in a comment c, and $\sum_k n_{k,j}$ is the total number of terms in the document c. The inverse term frequency idf_i is defined as:

$$idf_i = \frac{|\mathcal{C}|}{|\mathcal{C} : t_i \in c|} \tag{3}$$

where $|\mathcal{C}|$ is the total number of comments and $|\mathcal{C} : t_i \in c|$ is the number of comments containing the term t_i.

As the terming schema we have employed two different alternatives. First, we used the word as term. Second, we have used a n-gram approach as terms. N-gram is the overlapping subsequence of n words from a given comment.

- **Number of references to the comment (in-degree):** It indicates the number of times the comment has been referenced in other comments of the same news story. In 'Menéame' the reference is indicated by the symbol '#' followed by the comment number. This measure should be effective in capturing the importance of a comment in the whole discussion.

- **Number of references from the comment (out-degree):** It indicates the number of references of the comment to other comments of the same news story. We consider that this feature captures if the comment is talking about the news story or, instead, is a comment about other comment.

- **The number of the comment:** We also use the number of the comment which indicates the oldness of the comment. In 'Menéame', as happens also in other media, if a news story has a high number of comments, the main topic has usually derived to a discussion which, also, may be controversial.

- **The similarity of the comment with the snippet of the news story:** We used the similarity of the VSM of the comment with the model of the snippet of the news story. In particular, we employ the cosine similarity (Tata & Patel, 2007):

$$sim(\vec{v}, \vec{u}) = \cos(\theta) = \frac{\vec{v} \cdot \vec{u}}{||\vec{v}|| \cdot ||\vec{u}||} \tag{4}$$

where $\vec{v} \cdot \vec{u}$ is the inner product of $\vec{v}$ and $\vec{u}$ whereas $||\vec{v}|| \cdot ||\vec{u}||$ is the cross product of $\vec{v}$ and $\vec{u}$.

This value ranges from 0 to 1, where 0 means that the two of them are completely different (i.e., the vectors are orthogonal between them) and 1 means that the executables are equivalent.

We have used this feature because it can indicate us how much the comment relates to the news story.

- **Number of coincidences between words in the comment and tags of the news story:** We have counted the number of words that appear in the comment and that are tags of the news story. We have used this measure because it should be indicative of how related the comment is respect to the news story.
- **Number of URLs in the comment body:** We have also counted the number of URLs within the comment body. This measure tries to indicate whether the comment uses external sources in order to support its asseveration, although it can also be a link to a funny picture.

2.2.2. Syntactic Features

In this category we count the number of words in the different syntactic categories. To this end, we performed a Part-of-Speech tagging using FreeLing[5]. The following features were used:

- Number of adjectives in the comment body.
- Number of numbers in the comment body.
- Number of dates in the comment body.
- Number of adverbs in the comment body.
- Number of conjunctions in the comment body.
- Number of pronouns in the comment body.
- Number of punctuation marks in the comment body.
- Number of interjections in the comment body.
- Number of determinants in the comment body.

[5]Available in `http://www.lsi.upc.edu/~nlp/freeling`

- Number of abbreviations in the comment body.
- Number of verbs in the comment body.

These features are intended to capture the user's type of language in a particular comment. For instance, a high-use of adjectives should be indicative of expressing an opinion. By capturing the type of language, the method may able to identify the controversy-level of the comment as well as the type of information contained in the comment.

2.2.3. Opinion Features

This category refers to features that indicate opinion. Specifically, we used the following features:

- **Number of positive and negative words:** We have counted the number of words in the comment with a positive meaning and the number of words in the comment with a negative meaning. We employed an external opinion lexicon[6]. Since the words in that lexicon are in English and 'Menéame' is written in Spanish, we have translated them into Spanish.
- **Number of votes:** The number of positive votes of the comment. In 'Menéame' the votes are given by other users.
- ***Karma*:** The *karma* is computed by the website and represents how important is the comment based on the amount of positive and negative votes to that comment.

In this way, we have used two features that are external to 'Menéame': the number of positive and negative words; and the opinion features that 'Menéame' has already computed. The latter ones are the number of positive votes of that comment and the 'karma', which is a concept used in 'Menéame' to moderate comments. These features are devoted to categorise the comment in its level of controversy because they indicate the opinion of the 'Menéame' community about the comment and, also, the polarisation of the comment by means of the number of positive/negative words.

[6]Available in `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`

2.3. Machine-Learning Classifiers

Machine-learning is an active research area within *Artificial Intelligence* (AI) that focuses on the design and development of new algorithms that allow computers to reason and decide based on data (Bishop, 2006).

Machine-learning algorithms can commonly be divided into three different types depending on the training data: supervised learning, unsupervised learning and semi-supervised learning. For supervised algorithms, the training dataset must be labelled (e.g., the class of a comment) (Kotsiantis, 2007). Unsupervised learning algorithms try to determine how data are organised into different groups named clusters. Therefore, data do not need to be labelled (Kotsiantis & Pintelas, 2004). Finally, semi-supervised machine-learning algorithms use a mixture of both labelled and unlabelled data in order to build models, improving the accuracy of solely unsupervised methods (Chapelle et al., 2006).

Because comments can be properly labelled, we use supervised machine-learning; however, in the future, we would also like to test unsupervised and semi-supervised methods for automatic moderation of comments.

2.3.1. Bayesian Networks

Bayesian Networks (Pearl, 1982), which are based on the *Bayes Theorem*, are defined as graphical probabilistic models for multivariate analysis. Specifically, they are directed acyclic graphs that have an associated probability distribution function (Castillo et al., 1996). Nodes within the directed graph represent problem variables (they can be either a premise or a conclusion) and the edges represent conditional dependencies between such variables. Moreover, the probability function illustrates the strength of these relationships in the graph (Castillo et al., 1996).

The most important capability of Bayesian Networks is their ability to determine the probability that a certain hypothesis is true (e.g., the probability of a comment to be appropriate) given a historical dataset.

2.3.2. Decision Trees

Decision Tree classifiers are a type of machine-learning classifiers that are graphically represented as trees. Internal nodes represent conditions regarding the variables of a problem, whereas final nodes or leaves represent the ultimate decision of the algorithm (Quinlan, 1986).

Different training methods are typically used for learning the graph structure of these models from a labelled dataset. We use *Random Forest*, an

ensemble (i.e., combination of weak classifiers) of different randomly-built decision trees (Breiman, 2001), and *J48*, the WEKA (Garner, 1995) implementation of the *C4.5* algorithm (Quinlan, 1993).

2.3.3. K-Nearest Neighbour

The *K-Nearest Neighbour* (KNN) (Fix & Hodges, 1952) classifier is one of the simplest supervised machine-learning models. This method classifies an unknown specimen based on the class of the instances closest to it in the training space by measuring the distance between the training instances and the unknown instance.

Even though several methods to choose the class of the unknown sample exist, the most common technique is to simply classify the unknown instance as the most common class amongst the K-nearest neighbours.

2.3.4. Support Vector Machines (SVM)

SVM algorithms divide the n-dimensional space representation of the data into two regions using a *hyperplane*. This hyperplane always maximises the *margin* between those two regions or classes. The margin is defined by the farthest distance between the examples of the two classes and computed based on the distance between the closest instances of both classes, which are called *supporting vectors* (Vapnik, 2000).

Instead of using linear hyperplanes, it is common to use the so-called *kernel functions*. These kernel functions lead to non-linear classification surfaces, such as polynomial, radial or sigmoid surfaces (Amari & Wu, 1999).

3. Empirical Validation

In this section we describe the validation we have conducted in order to test the suitability of our method.

3.1. Dataset creation

To comprise a dataset of comments from 'Menéame' we programmed an application that gather the comments of the news in the front page of 'Menéame' daily. This program was scheduled to run every day at 10:00 AM. In this way, we retrieved a dataset with the news from 5th of April, 2011 to 12th of April, 2011, comprising a week of news. This dataset had a total number of 9,044 comments.

Thenceforth, we labelled each of the comments in their three categories[7]:

- **Focus of the comment:** It can be focused on the news story or it can be focused on other comment.
- **Type of information:** It can be a contribution, irrelevant to the news story or an opinion.
- **Controversy level:** It can be normal, controversial, very controversial or a joke.

Table 1: Distribution of samples for the categorisation about focus of the comment.

Class	Number of Comments
Focus on the news story	5191
Focus on other Comment	3853

Table 2: Distribution of samples for the categorisation about the type of information of the comment.

Class	Number of Comments
Contribution	217
Opinion	2460
Irrelevant	6367

Table 3: Distribution of samples for the categorisation about the controversy level of the comment.

Class	Number of Comments
Normal	6327
Controversial	1124
Very Controversial	1063
Joke	530

[7]The labelled dataset can be downloaded at `http://paginaspersonales.deusto.es/isantos/resources/Data_Meneame_dot_net_from_April_5th_to_April_12th.rar`

Table 1 shows the distribution for the first categorisation. In this way, most of the labelled comments focused on the news story. However, this class is not very unbalanced. Table 2 shows the distribution of the comments depending on its type of information. Most of the comments were irrelevant while only 217 were contributions to the information in the news story. Finally, Table 3 shows the number of comments in the classes regarding the controversy level of the comment.

3.2. Methodology

Using the labelled dataset, we coded an application to extract all the features described in Section 2. In order to build the VSM of the comment body two different approaches were employed: we computed the VSM with words as terms and, also, we used n-grams of a maximum value of $n = 3$ and a minimum value of $n = 1$. The two different models were computed to test whether n-grams can enhance the proposed approach or not. In order to build the model, we started by removing stop-words (Wilbur & Sirotkin, 1992), which are words devoid of content (e.g., 'a','the','is'). These words do not provide any semantic information and add noise to the model (Salton & McGill, 1983). We used an external stop-word list of Spanish words[8].

We extracted the top features for each of the classification types using *Information Gain* (Kent, 1983), an algorithm that evaluates the relevance of an attribute by measuring the information gain with respect to the class:

$$IG(j) = \frac{\sum_{v_j \in R} \sum_{C_i} P(v_j, C_i) \cdot (P(v_j, C_i)}{P(v_j) \cdot P(C_i)} \tag{5}$$

where C_i is the i-th class, v_j is the value of the j^{th} attribute, $P(v_j, C_i)$ is the probability that the j^{th} attribute has the value v_j in the class C_i, $P(v_j)$ is the probability that the j^{th} attribute has the value v_j in the training data, and $P(C_i)$ is the probability of the training dataset belonging to the class C_i. Using this measure, we removed any features that had a IG value of zero. In this way, we constructed six *ARFF* files (Holmes et al., 1994) (i.e., Attribute Relation File Format) with the resultant vector representations of the comments to build the aforementioned WEKA's classifiers: three for each

[8]The list of stop words can be downloaded at `http://paginaspersonales.deusto.es/isantos/resources/stopwords.txt`

type of VSM and one for each one of the three types of categorisation. Table 4 shows the number of features selected for each classification and VSM.

Table 4: Number of features for each categorisation.

Categorisation	# Features with words	# Features with n-grams
Focus of the comment	648	1,209
Type of information	1,413	3,616
Controversy level	276	920

Next, we evaluated the precision of the method to categorise the comments. To this extent, by means of the dataset, we conducted the following methodology to evaluate the proposed method:

- **Cross validation:** This method is generally applied in machine-learning evaluation (Bishop, 1995). In our experiments, we performed a K-fold cross validation with $k = 10$. In this way, our dataset is 10 times split into 10 different sets of learning (90 % of the total dataset) and testing (10 % of the total data).

- **Learning the model:** For each fold, we accomplished the learning step of each algorithm using different parameters or learning algorithms depending on the specific model. The algorithms use the default parameters in the well-known machine-learning tool WEKA (Garner, 1995). In particular, we used the following models:

  - *Bayesian networks (BN):* With regards to Bayesian networks, we utilise different structural learning algorithms: K2 (Cooper & Herskovits, 1991) and Tree Augmented Naïve (TAN) (Geiger et al., 1997). Moreover, we also performed experiments with a Naïve Bayes Classifier (Bishop, 1995).
  - *Support Vector Machines (SVM):* We performed experiments with a polynomial kernel (Amari & Wu, 1999), a normalised polynomial Kernel (Maji et al., 2008), a Pearson VII function-based universal kernel (Üstün et al., 2007) and a radial basis function (RBF) based kernel (Cho et al., 2008).
  - *K-nearest neighbour (KNN):* We performed experiments with $k = 1$, $k = 2$, $k = 3$, $k = 4$, and $k = 5$.

– *Decision Trees (DT):* We performed experiments with J48(the *Weka* (Garner, 1995) implementation of the *C4.5* algorithm (Quinlan, 1993)) and Random Forest (Breiman, 2001), an ensemble of randomly constructed decision trees. In particular, we tested random forest with a variable number of random trees N, $N = 10$, $N = 50$ and $N = 100$.

- **Testing the models:** To test the approach, we measured the Area Under the ROC Curve (AUC).

3.3. Results

Table 5: Results in terms of AUC of the categorisation about the focus of the comment.

Dataset	Word VSM	N-gram VSM
KNN K=1	0.83	0.76
KNN K=2	0.87	0.81
KNN K=3	0.89	0.84
KNN K=4	0.91	0.86
KNN K=5	0.92	0.87
Bayesian Network with K2 as structural learning algorithm	0.97	0.95
Bayesian Network with TAN as structural learning algorithm	0.99	0.99
Naive Bayes	0.83	0.69
SVM with Polynomial Kernel	0.96	0.96
SVM with Normalised Polynomial Kernel	0.97	0.96
SVM with Pearson VII function-based universal kernel	0.96	0.95
SVM with RBF kernel	0.95	0.95
Decision Tree: J48	0.97	0.96
Decision Tree: RandomForest N=10	0.99	0.99
Decision Tree: RandomForest N=50	0.99	0.99
Decision Tree: RandomForest N=100	0.99	0.99

Table 5 shows the results of the categorisation about the focus of the comment. In general, the results were better when the tokens used in the content body were words than when using n-grams. This categorisation was the easiest one and the results show that Random Forest achieved a 0.99 of AUC. Table 6 shows the results of the categorisation about the type of information of the comment. The results were also better when the tokens used in the content body were words instead of n-grams. However, the best results were obtained by Random Forest with 50 and 100 trees when the content body was tokenised with n-grams: a 0.85 of AUC. Table 7 shows

Table 6: Results in terms of AUC of the categorisation about the type of information.

Dataset	Word VSM	N-gram VSM
KNN K=1	0.65	0.59
KNN K=2	0.70	0.61
KNN K=3	0.72	0.61
KNN K=4	0.73	0.61
KNN K=5	0.74	0.61
Bayesian Network with K2 as structural learning algorithm	0.82	0.83
Bayesian Network with TAN as structural learning algorithm	0.82	0.83
Naive Bayes	0.78	0.68
SVM with Polynomial Kernel	0.77	0.80
SVM with Normalised Polynomial Kernel	0.73	0.73
SVM with Pearson VII function-based universal kernel	0.75	0.77
SVM with RBF kernel	0.51	0.51
Decision Tree: J48	0.72	0.74
Decision Tree: RandomForest N=10	0.81	0.82
Decision Tree: RandomForest N=50	0.83	0.85
Decision Tree: RandomForest N=100	0.84	0.85

the results of the categorisation about the controversy level. In this case the results were better using n-grams. The best results were obtained by Random Forest with 100 trees: a 0.70 of AUC. This categorisation is the hardest one because it is highly subjective.

These results show also which classification tasks are harder to perform. In this way, the results confirm that the easiest one is to find out if a comment focuses in the news story or in another comment whilst the hardest one is to realise the controversy level of the comment. They also raise the doubt about the subjectivity of the labelling, specially for the last categorisation level, regarding the different levels of controversy: they highly depend on the opinions of the person who labels the comments.

However, the results for three different types of categorisation are sound and our approach can be used to properly categorise each comment. In this way, the results also show that the n-gram approach is better in order to conform the VSM-based features than the classic words approach. Besides, the information about the comments can also be used to categorise users with regards to the type of comments they usually perform.

Table 7: Results in terms of AUC of the categorisation about controversy level of the comment.

Dataset	Word VSM	N-gram VSM
KNN K=1	0.59	0.61
KNN K=2	0.62	0.63
KNN K=3	0.63	0.64
KNN K=4	0.64	0.64
KNN K=5	0.64	0.64
Bayesian Network with K2 as structural learning algorithm	0.61	0.62
Bayesian Network with TAN as structural learning algorithm	0.64	0.66
Naive Bayes	0.63	0.63
SVM with Polynomial Kernel	0.56	0.57
SVM with Normalised Polynomial Kernel	0.59	0.58
SVM with Pearson VII function-based universal kernel	0.57	0.54
SVM with RBF kernel	0.50	0.51
Decision Tree: J48	0.58	0.63
Decision Tree: RandomForest N=10	0.62	0.64
Decision Tree: RandomForest N=50	0.66	0.69
Decision Tree: RandomForest N=100	0.67	0.70

4. Discussion and Further Work

The proposed method has been validated using data from 'Menéame'. The approach can categorise the comments made by users in three different ways. This method may be employed by administrators of webpages in order to moderate their website. For instance, it can be used to adequate the comments and visualisation of the page regarding the viewer, filter content that may damage the brand image of the page and also to categorise the users via their comments. Besides, despite we have focused on a social news site, such an approach can be adaptable to any webpage that allows its users to generate content for it.

However, although the obtained results confirm that our method is valid to accurately classify comments and moderate them, there are several topics of discussion in which we will focus in future versions of this system.

The use of supervised machine-learning algorithms for the model training, can be a problem in itself. In our experiments, we used a training dataset of one week. As the dataset size grows, so does the issue of scalability. This problem produces excessive storage requirements, increases time complexity and impairs the general accuracy of the models (Cano et al., 2006). To reduce disproportionate storage and time costs, it is necessary to reduce the

original training set (Czarnowski & Jedrzejowicz, 2006). In order to solve this issue, *data reduction* is normally considered an appropriate preprocessing optimisation technique (Pyle, 1999; Tsang et al., 2003). Such techniques have many potential advantages such as reducing measurement, storage and transmission; decreasing training and testing times; confronting the *curse of dimensionality* to improve prediction performance in terms of speed, accuracy and simplicity and facilitating data visualization and understanding (Torkkola, 2003; Dash & Liu, 2003). Data reduction can be implemented in two ways. On the one hand, *Instance Selection* (IS) seeks to reduce the evidences (i.e., number of rows) in the training set by selecting the most relevant instances or re-sampling new ones (Liu & Motoda, 2001). On the other hand, *Feature Selection* (FS) decreases the number of attributes or features (i.e., columns) in the training set (Liu & Motoda, 2008). We applied FS in our experiments when selecting the attributes with more than zero of Information Gain. Because both IS and FS are very effective at reducing the size of the training set and helping to filtrate and clean noisy data, thereby improving the accuracy of machine-learning classifiers (Blum & Langley, 1997; Derrac et al., 2009), we strongly encourage the use of these methods.

Besides, the dataset was not balanced for the different classes. To address unbalanced data, we can apply Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002), which is a combination of over-sampling the less populated classes and under-sampling the more populated ones. The over-sampling is performed by creating synthetic minority class examples. In this way, instances were still unique and classes became more balanced. Another possibility is to use cost-sensitivity learning. Cost-sensitive learning is a machine-learning technique where one can specify the cost of each error and the classifiers are trained taking into account that consideration (Elkan, 2001).

There is an issue derived from Natural Language Processing (NLP) when dealing with semantics: *Word Sense Disambiguation* (WSD). A troll user may evade our method by explicitly exchanging the key words of the comment with other polyseme terms and thus avoid detection. Thereby, WSD is considered necessary to perform most natural language processing tasks (Ide & Véronis, 1998). Hence, we propose the study of different WSD techniques (a survey of different WSD techniques can be found in (Navigli, 2009)) able to provide a semantics-aware moderation tool. However, such a semantic approach for moderation should have to deal with the semantics of different languages (Bates & Weischedel, 1993) and, therefore, be language dependant.

Our technique has several limitations due to the representation of comments. For instance, in the context of spam filtering, most of the filtering techniques are based on the frequencies with which terms appear within messages and spammers have started modifying their techniques to evade such filters. These techniques can be applied by a troll user of a social news website. For example, *Good Word Attack* is a method that modifies the term statistics by appending a set of words that are characteristic of legitimate, thereby bypassing filters. Nevertheless, we can adopt some of the methods that have been proposed in order to improve spam filtering, such as *Multiple Instance Learning* (MIL) (Dietterich et al., 1997). MIL divides an instance or a vector in the traditional supervised learning methods into several sub-instances and classifies the original vector based on the sub-instances (Maron & Lozano-Pérez, 1998). Zhou et al. (Zhou et al., 2007) proposed the adoption of multiple instance learning for spam filtering by dividing an e-mail into a bag of multiple segments and classifying it as spam if at least one instance in the corresponding bag was spam. We can adapt this approach to the our comment moderation tool. Another attack, known as *tokenisation*, works against the feature selection of the comment by splitting or modifying key message features, which renders the term representation as no longer feasible (Wittel & Wu, 2004). All of these attacks, which spammers have been adopting, should be taken into account in the construction of future filtering or moderation systems.

References

Amari, S., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*, 783–789.

Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Bates, M., & Weischedel, R. (1993). *Challenges in natural language processing*. Cambridge Univ Pr.

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer New York.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, *97*, 245–271.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cano, J., Herrera, F., & Lozano, M. (2006). On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining. *Applied Soft Computing Journal*, *6*, 323–332.

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1996). *Expert Systems and Probabilistic Network Models*. (Erste ed.). New York, NY, USA.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT Press.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cho, B., Yu, H., Lee, J., Chee, Y., Kim, I., & Kim, S. (2008). Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Transactions on Information Technology in Biomedicine*, *12*, 247–256.

Cooper, G. F., & Herskovits, E. (1991). A bayesian method for constructing bayesian belief networks from databases. In *Proceedings of the 1991 conference on Uncertainty in artificial intelligence*.

Czarnowski, I., & Jedrzejowicz, P. (2006). Instance reduction approach to machine learning and multi-database mining. In *Proceedings of the 2006 Scientific Session organized during XXI Fall Meeting of the Polish Information Processing Society, Informatica, ANNALES Universitatis Mariae Curie-Skłodowska, Lublin* (pp. 60–71).

Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial Intelligence*, *151*, 155–176.

Derrac, J., Garcıa, S., & Herrera, F. (2009). A First Study on the Use of Coevolutionary Algorithms for Instance and Feature Selection. In *Proceedings of the 2009 International Conference on Hybrid Artificial Intelligence Systems* (pp. 557–564). Springer.

Dietterich, T., Lathrop, R., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*, 31–71.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 2001 International Joint Conference on Artificial Intelligence* (pp. 973–978).

Fix, E., & Hodges, J. L. (1952). Discriminatory analysis: Nonparametric discrimination: Small sample performance. *Technical Report Project 21-49-004, Report Number 11*, .

Garner, S. (1995). Weka: The Waikato environment for knowledge analysis. In *Proceedings of the 1995 New Zealand Computer Science Research Students Conference* (pp. 57–64).

Geiger, D., Goldszmidt, M., Provan, G., Langley, P., & Smyth, P. (1997). Bayesian network classifiers. In *Machine Learning* (pp. 131–163).

Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: a machine learning workbench. (pp. 357–361).

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, *24*, 2–40.

Jindal, N., & Liu, B. (2007). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1189–1190). ACM.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining* (pp. 219–230). ACM.

Kent, J. (1983). Information gain and a general measure of correlation. *Biometrika*, *70*, 163–173.

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. In *Proceeding of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies* (pp. 3–24).

Kotsiantis, S., & Pintelas, P. (2004). Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*, *1*, 73–81.

Lerman, K. (2007). User participation in social media: Digg study. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops* (pp. 255–258). IEEE Computer Society.

Liu, H., & Motoda, H. (2001). *Instance selection and construction for data mining*. Kluwer Academic Pub.

Liu, H., & Motoda, H. (2008). *Computational methods of feature selection*. Chapman & Hall/CRC.

Maji, S., Berg, A., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8). IEEE.

Maron, O., & Lozano-Pérez, T. (1998). A framework for multiple-instance learning. *Advances in neural information processing systems*, (pp. 570–576).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, *41*, 10:1–10:69.

Pearl, J. (1982). Reverend bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 133–136).

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann.

Quinlan, J. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Quinlan, J. (1993). *C4. 5 programs for machine learning*. Morgan Kaufmann Publishers.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill New York.

Tata, S., & Patel, J. M. (2007). Estimating the Selectivity of tf-idf based Cosine Similarity Predicates. *ACM SIGMOD Record*, *36*, 75–80.

Torkkola, K. (2003). Feature extraction by non parametric mutual information maximization. *The Journal of Machine Learning Research*, *3*, 1415–1438.

Tsang, E., Yeung, D., & Wang, X. (2003). OFFSS: optimal fuzzy-valued feature subset selection. *IEEE transactions on fuzzy systems*, *11*, 202–213.

Üstün, B., Melssen, W., & Buydens, L. (2007). Visualisation and interpretation of support vector regression models. *Analytica chimica acta*, *595*, 299–309.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.

Wilbur, W., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of information science*, *18*, 45–55.

Wittel, G., & Wu, S. (2004). On attacking statistical spam filters. In *Proceedings of the 1st Conference on Email and Anti-Spam (CEAS)*.

Zhou, Y., Jorgensen, Z., & Inge, M. (2007). Combating Good Word Attacks on Statistical Spam Filters with Multiple Instance Learning. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence-Volume 02* (pp. 298–305). IEEE Computer Society.