Expert Systems with Applications 39 (2012) 13417–13425

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

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^{3}Lab), University of Deusto, Avenida de las Universidades 24, 48007 Bilbao, Spain

ARTICLE INFO

Keywords:
Spam detection
Information filtering
Content filtering
Machine-learning
Web categorisation

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 statistical, syntactic and opinion features and machine-learning classifiers to classify a comment within three 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.

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 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, 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 focusses 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.
- 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

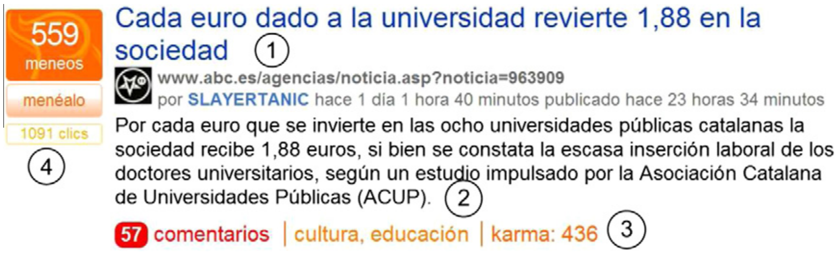
* Corresponding author. Tel.: +34 944139003; fax: +34 944139166.
E-mail addresses: isantos@deusto.es (I. Santos), jorge.delapenya@deusto.es (J. de-la-Peña-Sordo), iker.pastor@deusto.es (I. Pastor-López), patxigg@deusto.es (P. Galán-García), pablo.garcia.bringas@deusto.es (P.G. Bringas).

[1] http://digg.com/.

[2] http://meneame.net/.

[3] http://www.amazon.com/.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.061



559
meneos
menéalo
1091 clics

Cada euro dado a la universidad revierte 1,88 en la sociedad

www.abc.es/agencias/noticia.asp?noticia=963909
por SLAYERTANIC hace 1 día 1 hora 40 minutos publicado hace 23 horas 34 minutos

Por cada euro que se invierte en las ocho universidades públicas catalanas la sociedad recibe 1,88 euros, si bien se constata la escasa inserción laboral de los doctores universitarios, según un estudio impulsado por la Asociación Catalana de Universidades Públicas (ACUP).

57 comentarios | cultura, educación | karma: 436

Fig. 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.

Galli and Benjamín Villoslada and it is currently licensed as free software. At the beginning, it was focussed 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 four 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 two 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."*

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

Fig. 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.

Fig. 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 Fig. 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:
  - *Contribution:* The user contributes by adding new information. Fig. 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. Fig. 6 shows an irrelevant comment.
  - *Opinion:* These comments express the user's particular opinion about the topic discussed in the story. Fig. 7 shows an opinion.
- *Focus of the comment:* The comment can be focussed either on the main story or on another comment. Fig. 8 shows an example of a comment that focusses on the main story whilst Fig. 9 shows an example of a comment focussed on another comment. Although the comments that refer to another comment contain

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