A Web-Forum Free of Disguised Profanity by Means of Sequence Alignment[1]

Un foro web libre de obscenidades enmascaradas utilizando alineación de secuencias[2]

Christian Mogollón Pinzón[3]
Sergio Rojas-Galeano[4]

doi:10.11144/Javeriana.iyu20-2.wffd

How to cite this article:
C. Mogollón Pinzón and S. Rojas-Galeano, "A Web-forum free of disguised profanity by means of sequence alignment," Ing. Univ., vol. 20, no. 2, pp. 239-265, 2016. http://dx.doi.org/10.11144/Javeriana.iyu20-2.wffd

---

[1] Submitted on: November 20th, 2015. Accepted on: March 8th, 2016. This article is derived from the research project developed by the Laboratorio de Microelectrónica e Inteligencia Computacional (LAMIC) of the Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.

[2] Fecha de recepción: 20 de noviembre de 2015. Fecha de aceptación: 8 de marzo de 2016. Este artículo se deriva de un proyecto de investigación desarrollado por en el Laboratorio de Microelectrónica e Inteligencia Computacional (LAMIC) de la Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.

[3] Ingeniero de sistemas, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
E-mail: cgmogollonp@correo.udistrital.edu.co

[4] Ingeniero de sistemas, Universidad Nacional de Colombia. MSc. in Intelligent Systems, University College London (UCL). PhD. in Computer Science, UCL. Profesor asociado, Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. E-mail: srojas@udistrital.edu.co

Abstract

Profanity is the use of offensive, obscene, or abusive vocables or expressions in public conversations. A big source of conversations in text format nowadays are digital media such as forums, blogs, or social networks where malicious users are taking advantage of their ample worldwide coverage to disseminate undesired profanity aimed at insulting or denigrating opinions, names, or trademarks. Lexicon-based exact comparisons are the most common filters used to prevent such attacks in these media; however, ingenious users are disguising profanity using transliteration or masking of the original vocable while still conveying its intended semantic (e.g. by writing *piss* as *P!55* or *p.i.s.s*), hence defeating the filter. Recent approaches to this problem, inspired in the sequence alignment methods from comparative genomics in bioinformatics, have shown promise in unmasking such guises. Building upon those techniques we have developed an experimental Web forum (ForumForte) where user comments are cleaned of disguised profanity. In this paper we discuss briefly the techniques and main engineering artefacts obtained during the developing of the software. Empirical evidence reveals filtering effectiveness between 84% and 97% at vocable level depending on the length of the profanity (with more than four letters), and 86% at sentence level when tested in two sets of real user-generated-comments written in Spanish and Portuguese. These results suggest the suitability of the software as a language-independent tool.

Keywords

web forum; profanity detection; text analysis

Resumen

Por su carácter ofensivo o vulgar, las obscenidades son palabras o expresiones consideradas inapropiadas en conversaciones públicas. Hoy en día es común encontrar en *blogs*, foros y redes sociales el uso de obscenidades para insultar a o denigrar de opiniones, personajes o marcas; una anomalía cuyo agravante es mayor si se tiene en cuenta la amplia cobertura mundial que pueden alcanzar. El uso de diccionarios de palabras vetadas como mecanismo de filtrado es insuficiente, debido a la versatilidad del lenguaje escrito, que permite a los usuarios inventar variantes con transliteraciones o enmascaramientos del texto (por ejemplo, cambiar *mierda* por *m1erd@* o *m.i.e.r.d.a*). Inspirados en la genómica comparativa, se ha desarrollado un foro web experimental (ForumForte), donde los mensajes ingresados por los usuarios son inspeccionados y depurados de obscenidades transliteradas o enmascaradas. Este artículo presenta dicho *software* con una descripción breve de su diseño y su uso con datos reales de comentarios provenientes de medios digitales en español y portugués. La efectividad se ubicó entre 84 % y 97 % en la escala de palabra, dependiendo de la longitud de la obscenidad (para más de cuatro letras), y en 86 % en la escala de comentario. Estos resultados insinúan la utilidad del *software* para filtrado de obscenidades en foros web, independientes del idioma del usuario.

Palabras clave

foros web; detección de obscenidades; análisis de texto.

Introduction

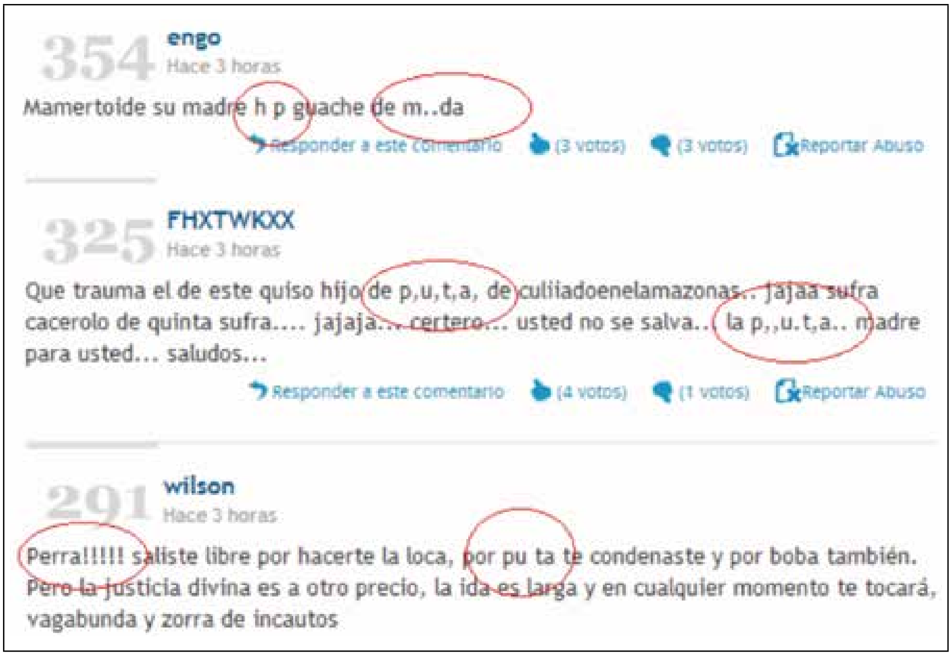
An essential feature of Web2.0 digital media is their ability to crowdsource user-generating-content, motivating collaboration and mutual construction of scenarios so as to yield richer user experiences [1]. An illustrative example is digital forums and blogs, where multiple users generate written comments about their own or other's opinions. Unfortunately, some users abuse of this freedom of speech for inappropriate purposes such as insulting, degrading, or boosting opinions, participants, brands or any other concept by means of offensive or obscene language. For these reasons, usually this kind of digital services must be moderated by website administrators in order to guarantee profanity–free user-generated text content.

Lexicon-based filters, which screen text against a blacklist of forbidden terms are a naïve moderation tool. A weakness in these filters is that they carry on exact comparisons missing variants with involuntary typos or misspellings, or more worryingly, variants disguised with transliterated or masking symbols written deliberately to circumvent the filter; the resulting variants still visually convey the actual meaning of the profanity.

Take for example the vulgar slang term *piss* transliterated as *P!55* or masked as *p-i-s-s* or worse still, a combination of both, *P-!-5-5*. Any of these attacks would easily defeat a literal comparison filter, but the message would still be clear for most readers. It is evident that the number of guises of this type grows combinatorially in size; thus, the lexicon-based approach is impractical. The anomaly is illustrated in Figure 1.

Similar anomalies have been identified in many other digital platforms; [2], [3] and even more, have been characterized as a security threat [4]. Recent approaches to tackle this problem taking inspiration on bioinformatics techniques for sequence alignment of genomes from different organisms have shown promising results [5], [6]. Building upon those results, we have developed an experimental profanity–safe Web forum (ForumForte). Our software was conceived as a concrete application of our previous results on the problems of revealing masked terms in spam email [5], automatic evaluation of fill-in-the-blank questionnaires [7], and automatic syntax verification of short blocks of code in programming languages [8]. An in-depth technical description of the filtering mechanism would be reported in a forthcoming paper.

Figure 1. Disguised profanity in a Colombian newspaper forum



354 engo
Hace 3 horas

Mamertoide su madre h p guache de m..da

Responder a este comentario (3 votos) (3 votos) Reportar Abuso

325 FHXTWKXX
Hace 3 horas

Que trauma el de este quiso hijo de p,u,t,a, de culiiadoenelamazonas.. jajaa sufra cacerolo de quinta sufra.... jajaja... certero... usted no se salva... la p,,u.t,a.. madre para usted... saludos...

Responder a este comentario (4 votos) (1 votos) Reportar Abuso

291 wilson
Hace 3 horas

Perra!!!!! saliste libre por hacerte la loca, por pu ta te condenaste y por boba también. Pero la justicia divina es a otro precio, la ida es larga y en cualquier momento te tocará, vagabunda y zorra de incautos

Source: author's own presentation of user-generated-content examples gathered from news forum from eltiempo.com

In the following sections we describe the software, the technology behind it, and a proof-of-concept of its potential application for content moderation in mainstream applications such as newspaper forums and micro–blogging media platforms. This is an extended version of a short paper recently published in the *Proceedings of the 10th Colombian Computer Conference* (10CCC) that was held in Bogotá on September, 2015 (see [9]). We note in passing that ForumForte is distributed as free software under the New BSD License and is available online or for download at: http://tinyurl.com/ForumForte.

1. Materials and Methods

1.1. Similarity Trees of Disguised Profanity

As it was observed before, the lexicon-based (exact comparison) approach against profanity disguising is not practical. A similar comparison difficulty was faced by bioinformaticians some decades ago in the field of comparative genomics, where the goal was to find common genetic motifs between different families of species. The "text" in that case was regarded as the sequences of DNA and protein molecules from the genome and proteome of living organisms [10], [11], written in an alphabet of letters representing the initials of their molecules (in the genome case {A, G, C, T} for (A)denine, (C)ytosine, (G)uanine and (T)hymine). Different organisms would have different genomes but when the sequences are aligned, similarities between sub–regions (genes) are found, except for a few places that differ. The small variations are due to mutations that insert, delete, or substitute one molecule or the other. The mutation may imply a change on the function of the phenotype that the gene codes for.

An example of sequence alignment is shown in Figure 2, where a gene whose phenotype is expressed during the synthesis of vitamin C is depicted for six species of mammals. The resemblance of the sequences is striking, although some differences, most surely due to mutations, are highlighted. As it can be seen, the last exon of the gene is very similar among cows, dogs, and rats, whereas it differs in one molecule deleted across humans and great apes. It is known that the former group can make their own vitamin C whereas the latter group has to take it from diet.

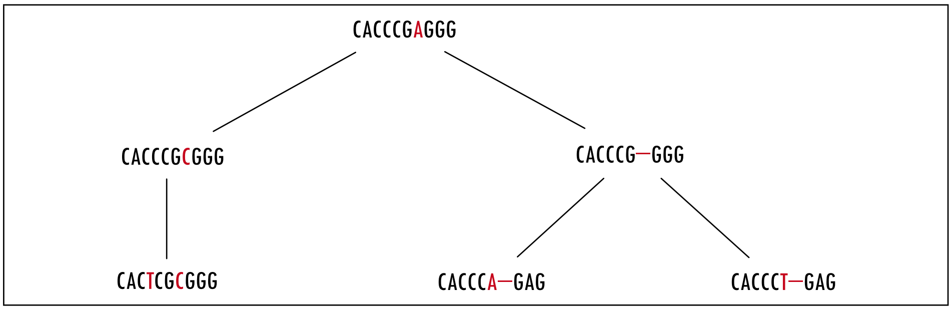
Figure 2. Sequence alignment of genes coding for vitamin C synthesis in six species

Human	ACCCTGAGGT	GGTGTCCCAC	TACCTGGTGG	GGGTACGCTT	CACCTG-GAG
Chimpanzee	ACCCCGAGGT	GGTGTCCCAC	TACCTGGTGG	GGCTACGCTT	CACCTG-GAG
Orangutan	ACCCTGAGGT	GGTGTCCCAC	TACCCGGTGG	GGGTGGGCTT	CACCCA-GAG
Cow	GCCCCAAGGT	AGTGGCCCAC	TACCCCGTGG	AGGTACGCTT	CACTCGCGGG
Dog	ACCCCAAGAT	GGTGGCCCAC	TTCCCTGTGG	AGGTCCGCTT	CACCCGCGGG
Rat	ACCCCAAAGT	GGTAGCCCAC	TACCCCGTAG	AGGTGCGCTT	CACCCGAGGG

Source: [12].

The evolutionist assumption that explains these variations in the genome is that mutations occurred during millions of generations of descendants from a common ancient genome, yielding the diversification of different species. Such diversification can be depicted as a phylogenetic tree, where branches grow every time a preserved mutation happens. In the previous example, the tree depicting mutations in the last exon is shown in Figure 3 (mutations highlighted in red). The genetic code of the rat is placed as the root, meaning that it would be the common ancestor of the six species (variants). From there, two branches diverge: one branch goes through the dog (substitution, A → C) down to the cow (another substitution, different location, C → T); the other branch goes to an unknown middle ancestor (a deletion, C) from where two branches open up, one for the orangutan (substitution, G → A), and one for humans and chimpanzees (substitution, G → T).

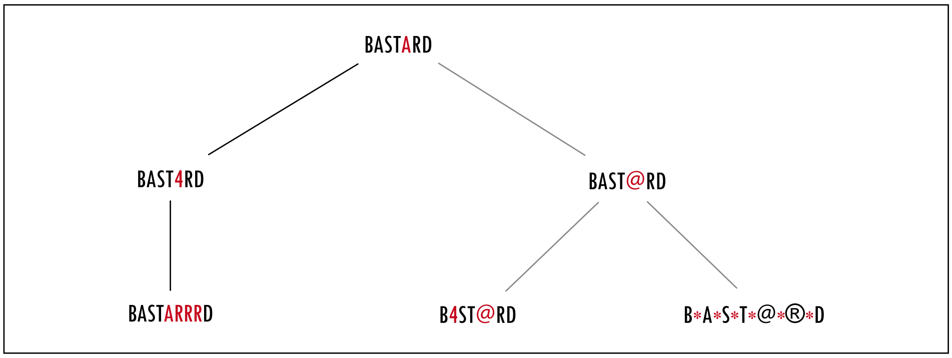
Figure 3. Phylogenetic tree of the gene exon coding for vitamin C in Figure 2



Source: authors' own elaboration

We adapted the idea of phylogenetic tree diversification to the profanity disguise anomaly described earlier. That is, we assume that the guises of a profanity vocable grow down in a similarity tree from a common ancestor (its canonical text) to the variants obtained by recurring application of edits or corrections made on the predecessors (see Figure 4, edits highlighted in red). Instead of keeping all the possible trees whose depth increase combinatorially with all possible edits, the idea behind our profanity detection mechanism is to trace back the disguised variant up to its common ancestor via classical sequence alignment algorithms [10], [11] or alternatively, using approximate string matching algorithms [13] which were independently developed by computer scientists for the same purpose. The key concept of these algorithms is the edit distance between two texts [14], which is the number of character corrections (substitutions, deletions, or insertions) that are needed to transform one text into the other. A special-purpose distance would be designed to correctly account for the edits that yield the profanity guises, as we shall describe later.

Figure 4. Similarity tree of disguised variants of the offensive word BASTARD



Source: authors' own elaboration

The approximate sequence matching algorithm (see e.g. [10]-[12]) carries out a pairwise comparison of the characters in the two sequences (the user–generated text and the canonical profanity vocable), while accumulating the number of edits (insertions, deletions, or substitutions) needed to transform one sequence into the other. The essential insight for detecting transliteration is to overlook the substitution of visually "twin" symbols (*e.g.* substituting 'o' by any of {0, °, ó, ò, ö, ô, Ø, θ, O}, see Figure 5), whereas for masking, the insight is to overlook the insertion of bogus segmentation characters such as { . , * , ~ , ¦ , – , _ , : , ; , ,""}. These couple of edits should add no value to the distance (or difference) between the two sequences, whereas edits such as deletion, insertion or any other substitution should count. Hence, the design of the edit distance between two symbols in their respective sequences is outlined in Figure 6, where *d* indicates if the edit counts or not in each of the cases mentioned above (this function was originally introduced in [5]).

Figure 5. An excerpt of the lists of twins substitutions

a	b	c	d	e	f	g	h	i	j	k	l	m	n	ñ	o	p	q	r	s	t	u	v	w	x	y	z	•
a	B	c	d	e	f	g	h	i	J	K	l	m	n	n	o	p	q	r	s	t	u	u	w	x	Y	z	,
A	b	C	D	E	F	G	H	I	j	k	L	M	N	N	O	P	Q	R	S	T	U	U	W	X	y	Z	.
4	8	(	ð	é	f	6	#	1		c	|		Ñ	Ñ	0	þ	9	®	5		ü	ü		*	ÿ	2	-
â	g	[		ê		9		!		C	C		ñ	ñ	Ø	Þ			$		û	û			Y̲		^
ä	v	{		ë				ï							Ö				§		ù	ù			ý		\
à	V	Ç		è				î							ö				z		Ü	Ü			Ý		=
Ä		ç		œ				ì							ô				Z		ú	ú					"

Source: authors' own elaboration

Figure 6. Edit distance function used in the filter mechanism

```
function $d$ = edit-distance($a, b$)
if isNull($b$) then
    $d = 1$;                              /* deletion */
else if isNull($a$) and notBogusSegmentator($b$) then
    $d = 1$;                              /* insertion */
else if isNull($a$) and isBogusSegmentator($b$) then
    $d = 0$;                         /* void insertion */
else if isValidSubstitute($b, a$) then
    $d = 0$;                     /* twin substitution */
else
    $d = 1$;                /* any other substitution */
```

Source: authors' own elaboration

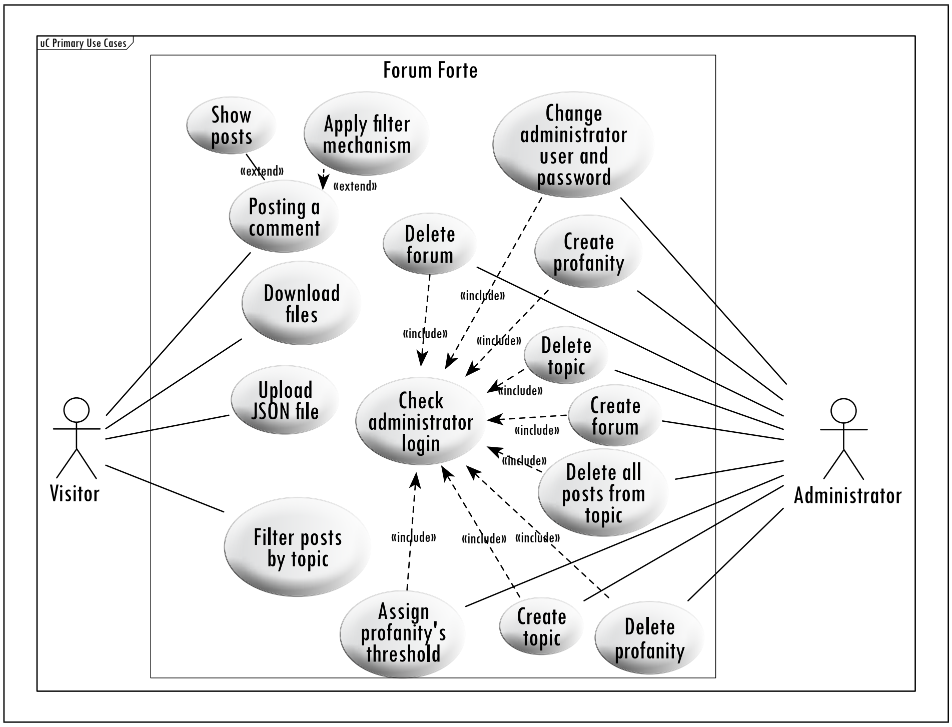
1.2. Software Design

We embarked on the development of ForumForte as a test-bed to verify the robustness of the filter mechanism described above, within an easy-to-use open forum where comments can be written about particular topics; no censorship is carried out, nor personal or usage information is collected. Simply put, Forum-Forte is a forum wall that screens comments against profanity; when profanity or disguised profanity is detected, the corresponding fragments are overwritten with a mask of asterisks, and both the original and filtered text sequences are posted to the wall. This software was designed as a Web application based on the software architectural MVC pattern [15] and the Java EE platform [16]. In the following, we shall describe the most representative design artifacts obtained during its development.

Let us start by summarizing the use–case scenarios for the software (Figure 7). There are two kinds of forum users: *visitors* and *administrator*. A *visitor* can inspect the fora pages as well as their contents, filter them by topics and indeed, post a comment in which case the profanity filter mechanism is activated and detection statistics would be collected. Lastly, he can download files for installation and upload a JSON file for batch processing of comments, as we will explain later on.

The other type of user is the *administrator*. This user has a password–protected account which allows him to carry out basic maintenance tasks such as forum and subject creation, elimination, cleaning, password update, etc. His other usages are related to the filter mechanism: updating the profanity canonical lexicon, looking at the performance statistics per forum or profanity vocable, and fine-tune the profanity tolerance parameters. These tolerances refer to the maximum edit distance for which two text sequences can be considered as equivalent (a value $\tau \in \{0, 1, 2, 3\}$).

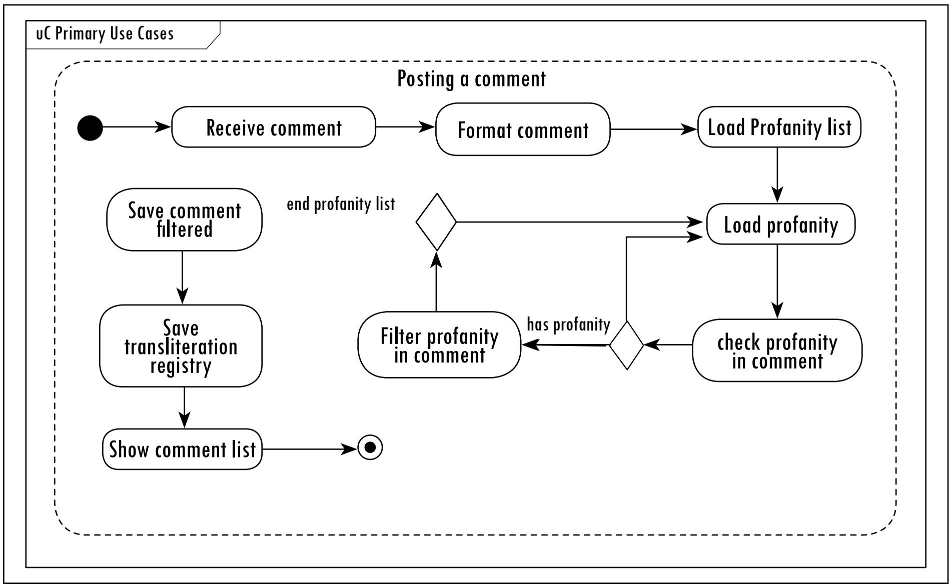
Figure 7. ForumForte use-case diagram



Source: authors' own elaboration

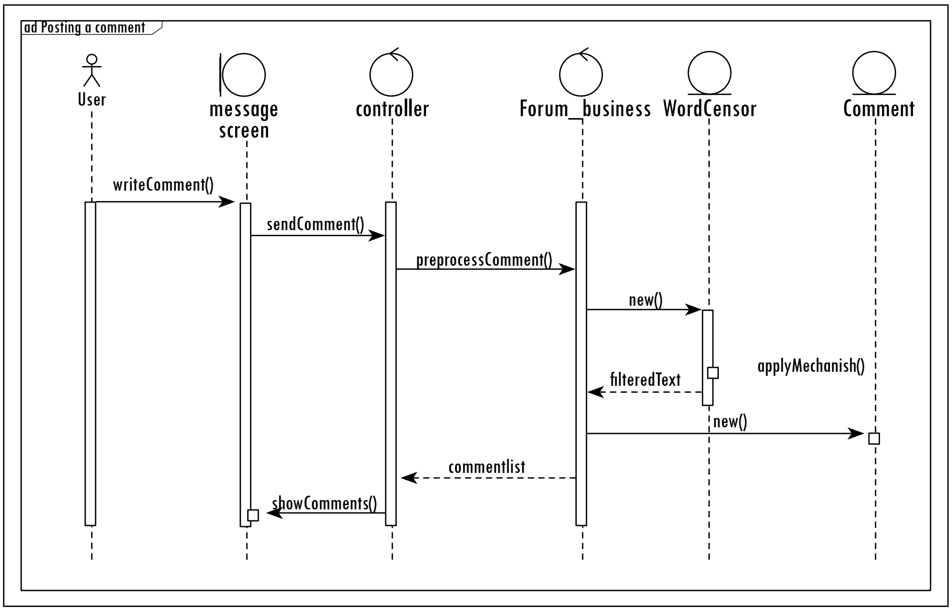
For illustration purposes, below we present the dynamical models of one of the main use cases in the forum, namely *posting a comment*. The activity diagram is shown in Figure 8, and the sequence diagram in Figure 9. All the other use-case dynamical models are available on request.

Figure 8. *Posting a comment* activity diagram



Source: authors' own elaboration

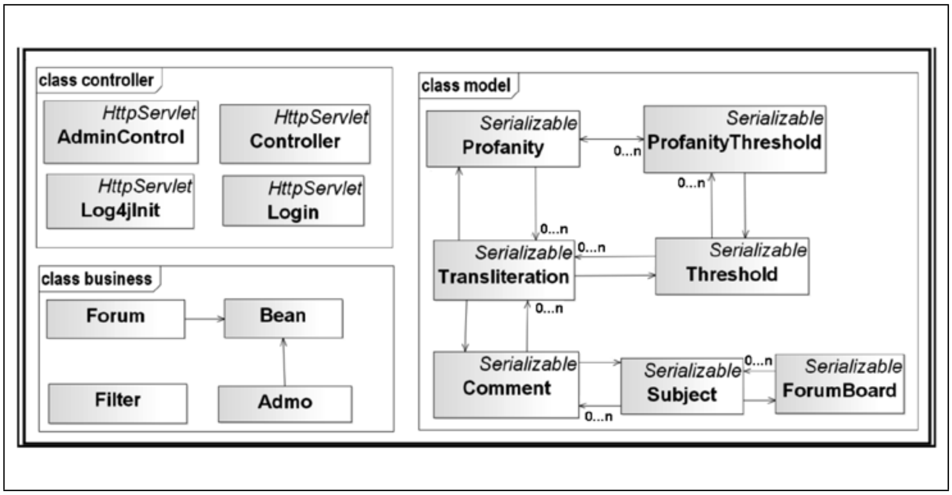
Figure 9. *Posting a comment* sequence diagram



Source: authors' own elaboration

The structural model of the software was designed as an MVC-based class diagram organized in three packages, namely business, model, and controller (see Figure 10). The business package includes classes Forum, Admin, Filter and Bean; these classes implement the logics of the functionalities previously described. On the other hand, the model package encapsulates ForumBoard, Subject, Threshold, Transliteration, Profanity; these classes are responsible of managing the visualization of forum comments and user interface. Finally, the controller package consists of classes Controller, Login, Log4jInit and Admin-Control; these classes control interaction with both kinds of users. Detailed views of this general model are also available.

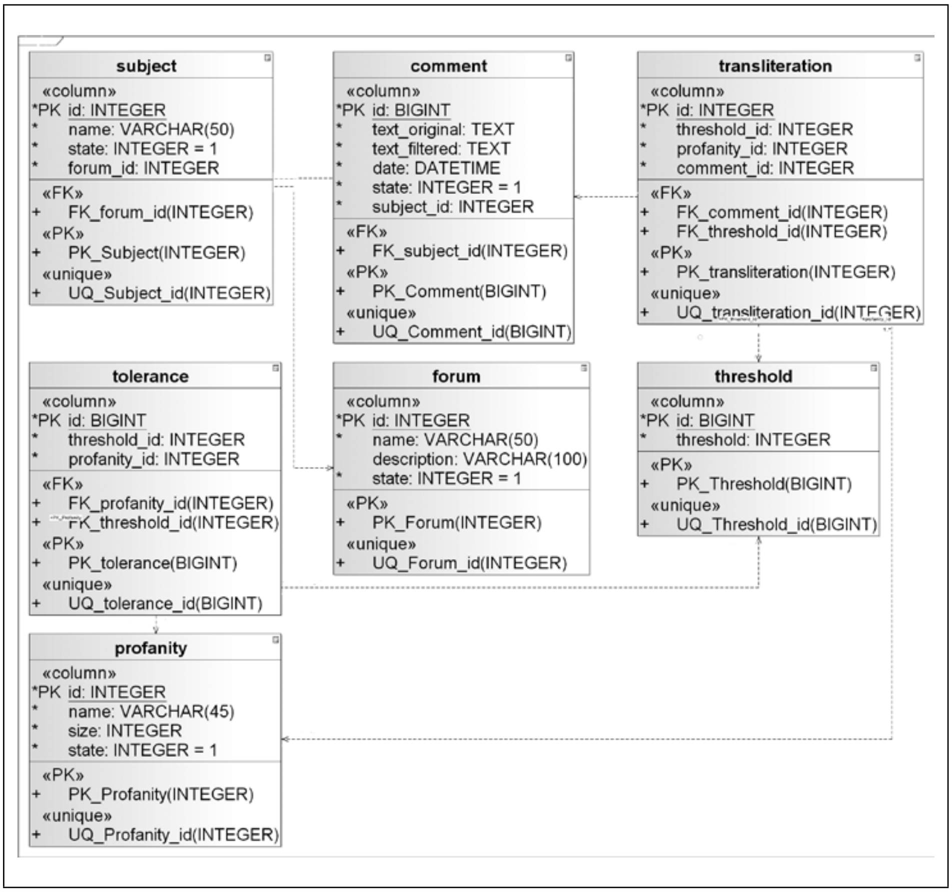
Figure 10. ForumForte class diagram overview



Source: authors' own elaboration

Next we discuss briefly the persistence model of the software, which was implemented as a relational database using the Java Persistence API framework that carries out the mapping from the structural model. The corresponding ER diagram is shown in Figure 11, consisting of tables: forum, subject, comment, filteredComment, profanity, and tolerance. The latter were included because the administrator may fine–tune the edit-distance tolerance per profanity term, and therefore statistics per tolerance are collected.

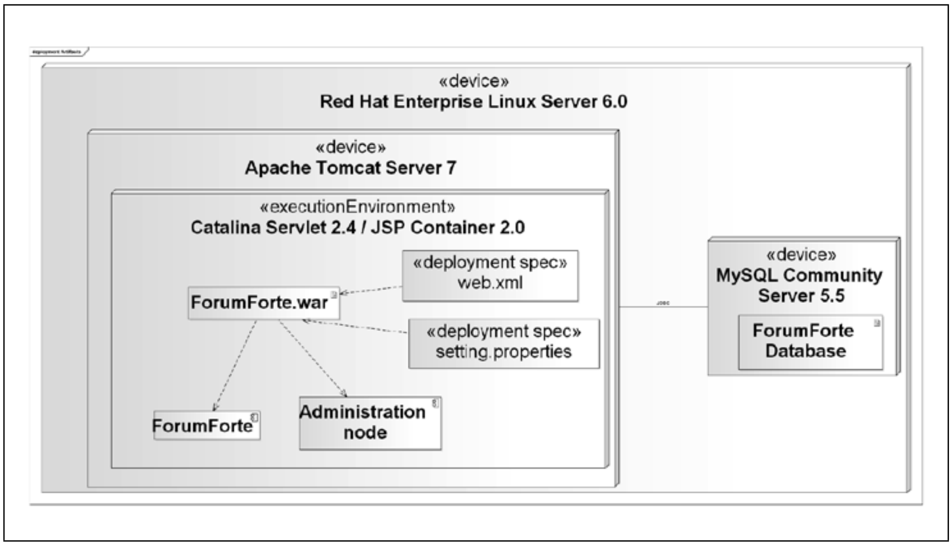
Figure 11. ForumForte persistence model



Source: authors' own elaboration

Finally, the deployment diagram of Figure 12 shows the sub-systems used to run the software as a Website application. It can be seen there that the Catalina Server 2.4 provides the execution environment for the components Administration and WebForum included in the ForumForte.war package. System configuration is defined in the corresponding properties and XML files. This environment runs on an Apache Tomcat 7.0.42 server within a Red Hat v6 Linux OS. Data persistence is achieved through a socket connection to a MySQL 5.5 server within a ForumForte database session.

Figure 12. ForumForte deployment diagram



Source: authors' own elaboration

1.3. Datasets

Two real-life datasets of user-generated-comments were fed to the software in order to test its effectiveness. The first dataset is a collection of 300 user comments (in Spanish) from the publicly–available news forums of a Colombian newspaper, traced during a time frame of 25 months (01-Jan-2011 to 31-Jan-2013). Every message in this dataset contains profanity that was not blocked by the forum filter.

A wide assortment of comments was included in this dataset: regarding word length, 58% (176) comments are shorter than 30 words, 25% (74) are medium-long, between 31-50 words, and the remainder 17% (50) ranges between 51 and 150 words. The majority of comments include only one use of profanity (76% or 227 comments), followed by two uses of profanity (20% or 61 comments); there are extreme cases such as a comment containing seven occurrences of swearing, and cases where the full comment is precisely one swearing term. The dataset contains 9293 words in total, 537 of which are swearing (5.8%).

Associated with this dataset, we also gathered a lexicon of 60 Spanish swearing terms, varying in length, the mode being 4-6 letters long (60%), with extremes being three 9 letter-long and one 2-letter long words. Excerpts of the datasets are illustrated in Table 1 and Table 2.

Table 1. Example of profanity comments from the Spanish dataset

De verdad g-u-e-v-o-n-e-s..?????… y cuanto consigno? … que m-a-r-i-c-a-d-as- -de noticias, este periodico ahora divierte es buscandole los p-u-to-s gazapos jejeje…!!!!
Quien le cree a estas malpar1das pirañas.
h pta…que susto.!
mmmmmierrrrrrrda! de ahora en adelante no conio en nadie y borracho menos!
Déjenlo,es su elección sexual y nuestra Constitución lo protege, decía mi abuela: que cada persona haga de su cu_lo un tambor, (cu_lo sin la rayita para evitar que me escondan el comentario)

Source: authors' own elaboration

Table 2. Example of plain profanity from the Spanish lexicon

guevon	cacorro	coño
puto	imbecil	pendejo
mierda	Culo	verga
hijueput	cabron	puta
chocha	marica	gonorrea
mamada	malparid	maldit
pipi	cagar	culo
tetas	hp	cabron

Source: authors' own elaboration

The second dataset consists of 2500 user-generated-comments written in Portuguese, extracted from a sports news website. This dataset was previously made publicly available in [17]. In there, human annotators inspected the dataset and identified 521 messages with some use of profanity, either disguised or not. However, as we shall highlight next, ForumForte was able to discover overlooked comments with occurrences of swearing. The blacklist lexicon for this dataset contains 40 base profanities in the Portuguese language. Recalling the ample differences between Portuguese and Spanish in written aspects such as the use of special diacritics (e.g. "*ã*," "*ç*," "*ê*") and digraphs (e.g. "*lh*," "*nh*"), this dataset may provide interesting information about the potential of the software in different language-speaking communities.

Table 3. Example of profanity comments from the Portuguese dataset

ó grandessíssimo filho da put4, mais uma pérola da literatura contemporânea, estás cada vez melhor, melhor mesmo só os broch3s da tua, ou então as intervenções do peclise/ argália.
ja nem coco consigo fazer!
tu nao és capaz de comer a puuuuta da mae dos teus filhos bastardos que tens là em casa , tem que ser os vizinhos e os amigos a faze-lo seu korno,encornado ,...ah ah ah ah ah
,,então levar no cú dói?????
godinho kabrao pede a demissao ,...seu koirao,...

Source: author's own presentation of data gathered from [17]

Table 4. Example of plain profanity from the Portuguese lexicon

bosta	cornos	panasca
broche	cuzinho	paneleiro
cabrão	cornudo	peida
cagalhão	cú	pila
cago	foder	picha
caralho	mamões	punhetas
chulecos	maricas	puta
cona	merda	rabeta

Source: author's own presentation of data gathered from [17]

2. Results

2.1. Software Execution

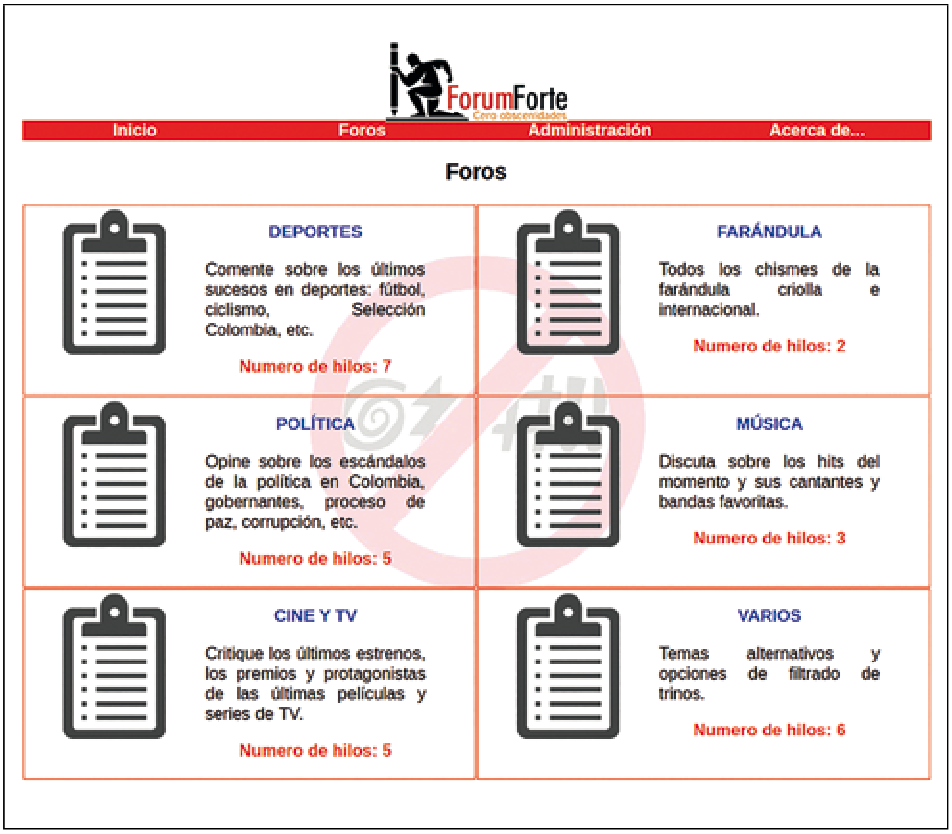
2.1.1. Online Release and On-Site Installation

ForumForte is installed an available online in http://tinyurl.com/ForumForte (last visit: April 15th, 2016). The Web application works on any Web browser (Firefox, Chrome, Explorer, and Safari). Alternatively, interested users can download and install the software in their own servers (user and installation guide are also available from the same URL).

2.1.2. Forum visitor usage

A visitor may browse the available fora by clicking the *Foros* option in the menu bar, which redirects to the page shown in Figure 13. Any choice here would display the list of subjects in each forum previously defined by the administrator. Then, by choosing one of the subjects, the visitor would be taken to the actual forum page, where comments made by other visitors would be shown (see Figure 14). We remark that interaction with the forum is anonymous and no private or network access information is collected by the system.

Figure 13. List of available fora



Foros

DEPORTES

Comente sobre los últimos sucesos en deportes: fútbol, ciclismo, Selección Colombia, etc.

Numero de hilos: 7

FARÁNDULA

Todos los chismes de la farándula criolla e internacional.

Numero de hilos: 2

POLÍTICA

Opine sobre los escándalos de la política en Colombia, gobernantes, proceso de paz, corrupción, etc.

Numero de hilos: 5

MÚSICA

Discuta sobre los hits del momento y sus cantantes y bandas favoritas.

Numero de hilos: 3

CINE Y TV

Critique los últimos estrenos, los premios y protagonistas de las últimas películas y series de TV.

Numero de hilos: 5

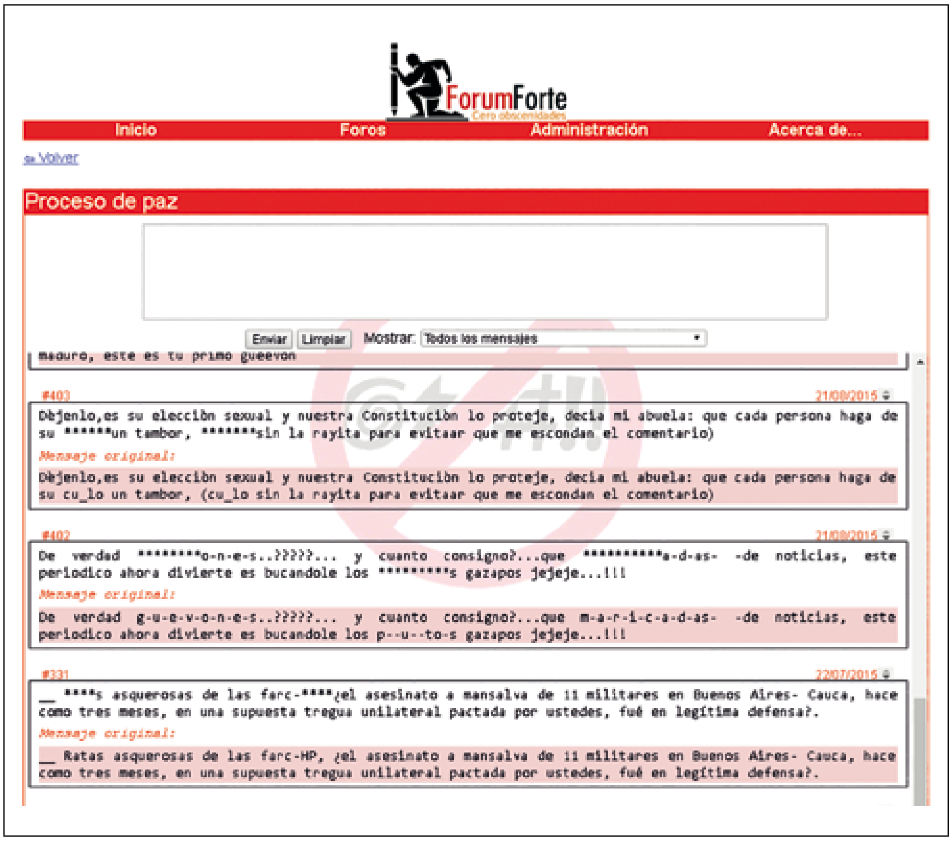
VARIOS

Temas alternativos y opciones de filtrado de trinos.

Numero de hilos: 6

Source: snapshot from ForumForte website

Figure 14. The appearance of an arbitrary forum



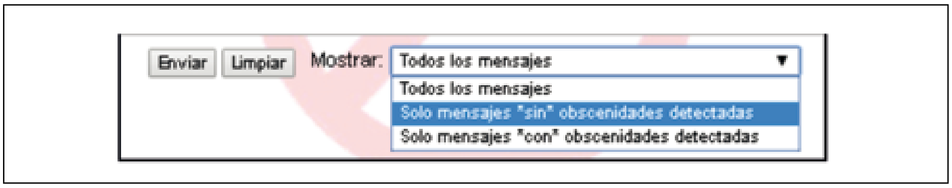
```
maduro, este es tu primo gueevon

#403                                                        21/08/2015
Dèjenlo,es su elecciòn sexual y nuestra Constituciòn lo proteje, decia mi abuela: que cada persona haga de
su ******un tambor, *******sin la rayita para evitaar que me escondan el comentario)
Mensaje original:
Dèjenlo,es su elecciòn sexual y nuestra Constituciòn lo proteje, decia mi abuela: que cada persona haga de
su cu_lo un tambor, (cu_lo sin la rayita para evitaar que me escondan el comentario)

#402                                                        21/08/2015
De  verdad  ********o-n-e-s..?????...  y  cuanto  consigno?...que  **********a-d-as-  -de  noticias,  este
periodico ahora divierte es bucandole los *********s gazapos jejeje...!!!
Mensaje original:
De  verdad  g-u-e-v-o-n-e-s..?????...  y  cuanto  consigno?...que  m-a-r-i-c-a-d-as-  -de  noticias,  este
periodico ahora divierte es bucandole los p--u--to-s gazapos jejeje...!!!

#331                                                        22/07/2015
__ ****s asquerosas de las farc-****¿el asesinato a mansalva de 11 militares en Buenos Aires- Cauca, hace
como tres meses, en una supuesta tregua unilateral pactada por ustedes, fuè en legìtima defensa?.
Mensaje original:
__ Ratas asquerosas de las farc-HP, ¿el asesinato a mansalva de 11 militares en Buenos Aires- Cauca, hace
como tres meses, en una supuesta tregua unilateral pactada por ustedes, fuè en legìtima defensa?.
```

Source: snapshot from ForumForte website

The layout of an actual forum is very intuitive and easy to use. Basically, there is a text box on the top of the page for the visitor to write his or her comment. The visitor can choose to either clean the current content of the text box, or to post the comment in which case, it would be screened by the filter engine. Once the text is processed, the original and filtered text would be posted to the forum wall as the most recent entry (just below the text box).

Each entry consists of the filtered text sequence aligned over the original text. If one or many disguised profanities are detected, they would be overwritten with a mask of asterisks on the top line of the entry. Comments are kept in the wall in chronological order, most recent first. Lastly, the visitor may choose to display all comments in the forum, only profanity–marked comments, or only profanity–free comments (see Figure 15).

Figure 15. Display options of forum comments

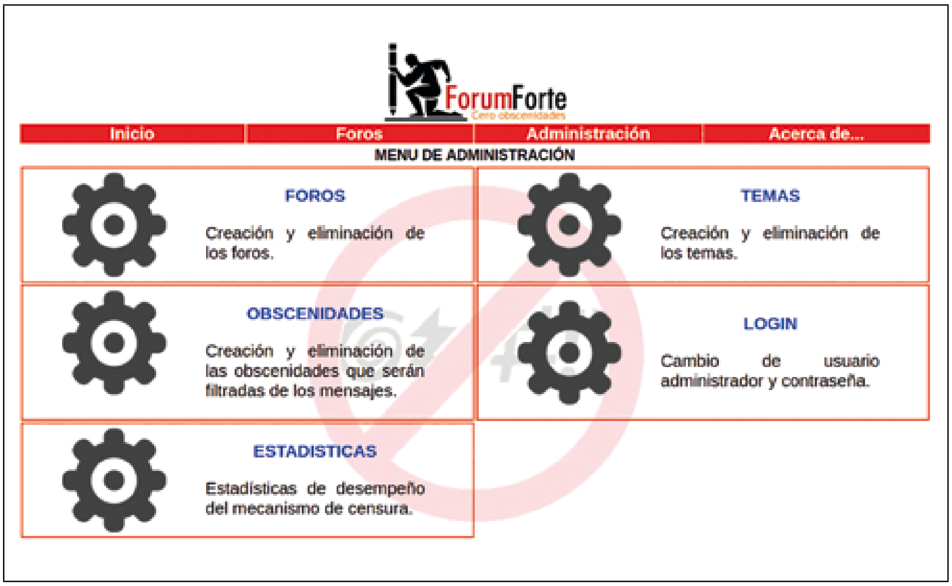

Enviar
Limpiar
Mostrar:
Todos los mensajes
Todos los mensajes
Solo mensajes "sin" obscenidades detectadas
Solo mensajes "con" obscenidades detectadas

Source: snapshot from ForumForte website

2.1.3. Forum administration

The administrator main page shows the dash board of Figure 16, where typical maintenance operations in forums, subjects, profanity lexicon, and statistics reports are carried out. For this purpose, the user should choose the *Administración* option in the menu page, and confirm his identity with a valid password. The software supports a unique administrator whose username and password can be updated at convenience using the *Login* choice. The remainder options aforementioned redirect to Web pages where the administrator can create, eliminate, or clean contents of the corresponding item.

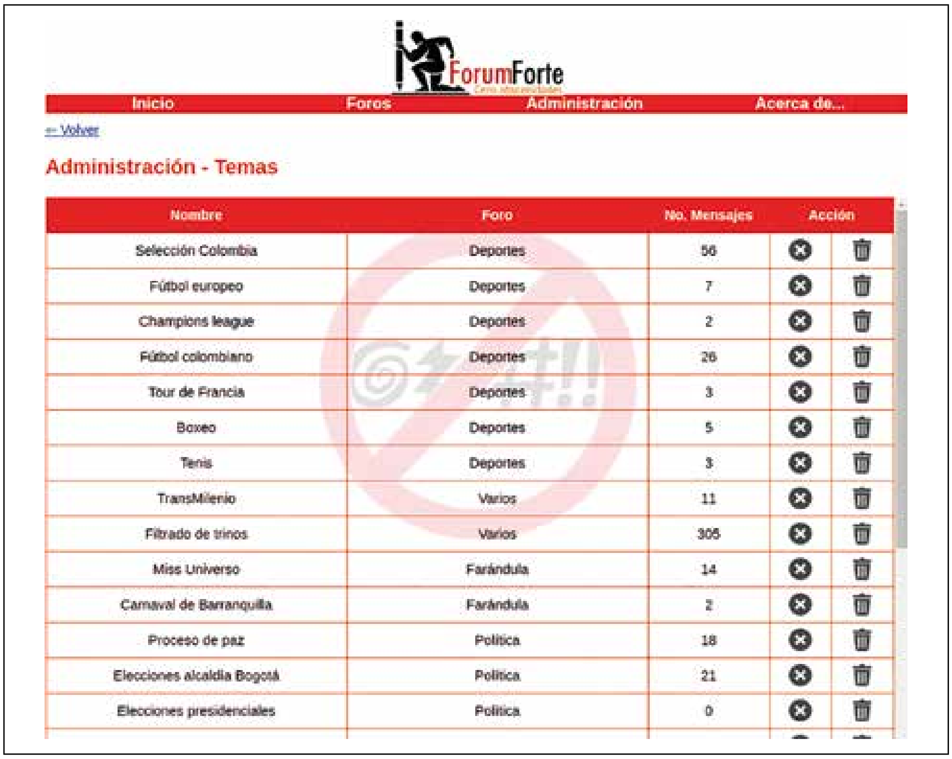
Figure 16. The administrator dash board


ForumForte
Cero obscenidades
Inicio
Foros
Administración
Acerca de...
MENU DE ADMINISTRACIÓN
FOROS
Creación y eliminación de los foros.
TEMAS
Creación y eliminación de los temas.
OBSCENIDADES
Creación y eliminación de las obscenidades que serán filtradas de los mensajes.
LOGIN
Cambio de usuario administrador y contraseña.
ESTADISTICAS
Estadísticas de desempeño del mecanismo de censura.

Source: snapshot from ForumForte website

For the sake of illustration, a sample subject administration page is shown in Figure 17 where the clean up ⊗ and remove 🗑commands are visible. The administrator can navigate to the actual forums pages and visualize the comments or participate in the discussion by clicking over the forum name. Similarly, the profanity lexicon module allows creating, removing, or tuning the transliteration tolerance of non-admitted vocables in the forum pages. In addition, the statistics module reports detection rates discriminated by forum or by profanity vocable (see Figure 18). The latter are furthermore broken down into individual rates per tolerance parameter, providing valuable information for tuning purposes with respect to particular vocables and transliteration attack patterns. We remark that forum statistics are computed instantly with its current comment contents whereas profanity statistics are historical, beginning at the moment they were created or assigned a different tolerance parameter. Statistics are lost when the associated item is removed.

Figure 17. A sample subject administrator page

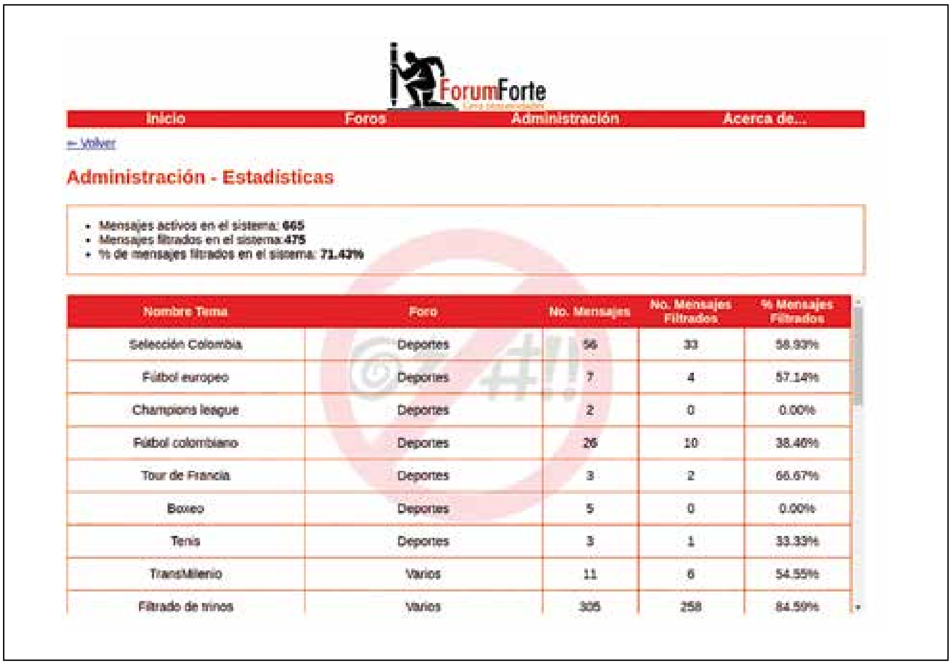


Administración - Temas

Nombre	Foro	No. Mensajes	Acción	
Selección Colombia	Deportes	56	⊗	🗑
Fútbol europeo	Deportes	7	⊗	🗑
Champions league	Deportes	2	⊗	🗑
Fútbol colombiano	Deportes	26	⊗	🗑
Tour de Francia	Deportes	3	⊗	🗑
Boxeo	Deportes	5	⊗	🗑
Tenis	Deportes	3	⊗	🗑
TransMilenio	Varios	11	⊗	🗑
Filtrado de trinos	Varios	305	⊗	🗑
Miss Universo	Farándula	14	⊗	🗑
Carnaval de Barranquilla	Farándula	2	⊗	🗑
Proceso de paz	Política	18	⊗	🗑
Elecciones alcaldía Bogotá	Política	21	⊗	🗑
Elecciones presidenciales	Política	0	⊗	🗑

Source: snapshot from ForumForte website

Figure 18. The performance statistics module



Administración - Estadisticas

- Mensajes activos en el sistema: **665**
- Mensajes filtrados en el sistema: **475**
- % de mensajes filtrados en el sistema: **71.43%**

Nombre Tema	Foro	No. Mensajes	No. Mensajes Filtrados	% Mensajes Filtrados
Selección Colombia	Deportes	56	33	58.93%
Fútbol europeo	Deportes	7	4	57.14%
Champions league	Deportes	2	0	0.00%
Fútbol colombiano	Deportes	26	10	38.46%
Tour de Francia	Deportes	3	2	66.67%
Boxeo	Deportes	5	0	0.00%
Tenis	Deportes	3	1	33.33%
TransMilenio	Varios	11	6	54.55%
Filtrado de trinos	Varios	305	258	84.59%

Source: snapshot from ForumForte website

ForumForte features an interesting interface to apply its filtering mechanism to external sources of user–generated text. This feature takes advantage of the JSON format for content extraction and storing provided by the Twitter® social network through its publicly–available API. This digital media platform is essentially a worldwide community forum for free short text messaging (comments no longer than 140 characters, known as tweets) with no moderation, and therefore a real–world practical scenario to test our development.

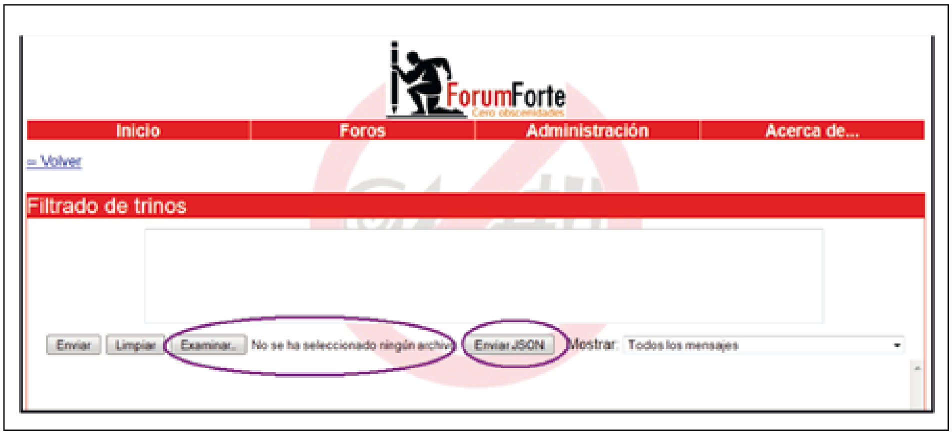
2.1.4. External input

In order to try this feature, the user should prepare an input file with a JSON format. Such file can be obtained by logging in into the Twitter API console with a valid user account[5] and then extracting tweets for a chosen user profile or trend topic. The file can then be processed in ForumForte entering the specially–designed forum found in the path *Foros→Filtrado de trinos*. In contrast with the other fora, this one features two buttons to load the JSON file and to submit

[5] The API Twitter is available at: https://dev.twitter.com/rest/tools/console (last visit: April 15th, 2016)

it to the filtering engine (see Figure 19), which would subsequently process each tweet in the file and post it to the forum wall as if it was originally typed in by a visitor. We highlight that by adhering to the Twitter JSON format, it is possible to filter user–generated text from other sources.

Figure 19. The specially designed forum for processing external JSON Twitter files



Source: snapshot from ForumForte website

2.2. Experiments with the Spanish Dataset

A first experiment aimed at testing the detection effectiveness of ForumForte was initially conducted on the Spanish dataset described in Section 2.3. Profanity ground truth was obtained by manually labelling the occurrences of disguised swearing from the profanity lexicon observed in each comment within the dataset. This experiment was carried out by processing the dataset using the JSON external input feature and statistics module of ForumForte. The 300 comments were screened against each profanity item in the lexicon, while verifying correct detections and logging statistics. Notice that the goal of this experiment was to evaluate detection rates at the vocable level, that is, correct identification of instances of swearing terms, either plain or corrupted, in the whole dataset.

The resulting detection rates are reported in Table 5, which demonstrate the feasibility and promise of the method in real world scenarios. For the sake of easier illustration of the results, they are presented summarized by grouping profanities according to their character length. In each group, the tests were repeated for the values $\tau = \{0, 1, 2, 3\}$ of the tolerance parameter. Grayed values in the table denote detection counters closest to the ground truth (values above the ground truth indicate occurrence of false positives).

Table 5. Detection rates in the real-life dataset

Obscenity length (m)	Ground truth	Detections per tolerance parameter			
		$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 3$
2	28	33	2652	720	720
3	33	25	2085	5774	1053
4	205	186	1920	12277	21449
5	136	115	281	2517	9502
6	33	32	60	461	4688
7	33	28	29	85	647
8	26	19	25	34	94
9	43	33	36	53	92

Source: authors' own elaboration

These results reveal that in profanity of short length ($m \leq 6$) a tolerance $\tau = 0$ is more effective. This may be explained because in short sequences, insertion or deletion of even just a single character would result in a not immediately easy to interpret variant of the original word (e.g. *coo*→*coño, pQuta*→*puta*); therefore attackers prefer to use substitutions to obtain profanity guises in these groups. On the other hand, medium or larger size profanity ($m \geq 7$), require higher tolerances $\tau = \{1, 2\}$ for better detection, since in these cases deletions and insertions are easier to interpret and thus attackers tend to use such edits rather than substitutions to disguise profanity. Lastly, a tolerance $\tau = 3$ is clearly not recommended in any group since it yields extremely high false positive rates.

In brief, we remark that the effectiveness of the software depends on the length of the profanity. For example, in lengths $m = 4, \ldots, 9$ the detection rates are: 91%, 85%, 97%, 88%, 96%, 84%. On the other hand, the rates for shorter profanities ($m = 2, 3$) are less confident (100% + false positives, 76%), which may be explained due to the difficulty to corrupt or disguise the original term with such limited number of letters without affecting its visual recognition (for example, the Spanish swearing term "*hp*" would become illegible with just one deletion or one non-twin substitution).

2.3. *Experiments with the Portuguese Dataset*

The second experiment was aimed at further testing suitability of the detection mechanism of ForumForte for other languages different to the one for which it was originally designed (Spanish). For this purpose, we conducted tests on the

Portuguese dataset described in Section 1.3. Experimental settings were analogous to those described in Section 2.2, except that in this case the goal was to evaluate detection effectiveness at the message level (because this dataset was originally characterized in that way [17]), that is, to correctly reject messages with any profanity content either plain or corrupted. Hence, since the experiment is a binary classification task, we report the results as a confusion matrix that is shown in Table 6. In order to simplify the analysis we set $\tau = 0$ for all the profanities in the lexicon.

Table 6. Confusion matrix for predictions on the Portuguese dataset

		ForumForte prediction		
		Profanity Positive	Profanity Negative	
Original labels	Profanity Positive	441 (85%)	80 (15%)	521
	Profanity Negative	85 (4%)	1894 (96%)	1979
		526	1974	2500

Source: authors' own elaboration

Let us first focus on the comments originally labeled as profanity. The software correctly rejected 85% (441 out of 521) of them and incorrectly accepted 15% (80 comments). A closer inspection on these 80 comments revealed that 20 of them contain variants of a two-letter long Portuguese swearing term (*cú*); again, this kind of really short profanities are harder to detect because the limited number of feasible edits to disguise them without losing their legibility, as it happened as well on the Spanish dataset experiment. Besides, without the accent, this particular sequence can be found as a syllable of many legitimate Portuguese words, thus an exact match would be better to achieve a more accurate detection. Notice that, provided that we overlook these 20 comments, the sensitivity of the software on this dataset would increase to 89%.

Now, regarding the comments originally labeled as non-profanity, the experiment obtained a false positive rate of 4% (85 comments). In this respect, again a closer inspection was conducted and we found that 40 of them were possibly mislabeled in the original dataset, that is, they actually may be instances of Portuguese profanity (an excerpt of these findings are shown in Table 7). Therefore, we modified these labels so that now the dataset contains 561 profanity comments; subsequently we run the experiments again and collected the results in the confusion matrix of Table 8. It can be seen that in the modified dataset the software improved its effectiveness, achieving a sensitivity of 86% and a false positive rate of 2%. This fact indicates that the software can also be used as a tool to corroborate annotation made by human moderators and even suggest overlooked cases for further inspection.

Table 7. An excerpt of profanity comments originally labeled as non-profanity

Id comment	Profanity [disguised=plain]	Contents
593	{lampibosta!=bosta}	lol a fazer a tipica festa! depois nao chores,, quero ver se no fim do campeonato estas aqui?! vais bater no namorado na sogra e no cao lampibosta! no ano passado tiveram a 5 pontos de vantagem ficaram a 6 do primeiro
736	{p.utas=puta}	que que saiba, o vitótó napeiad é que disse que o f.oi c.om p.utas ganhava sempre contra o benfica... se calhar foi por andar a cantar de galo que estava com tanta azia no final do jogo!
751	{pe i da=peida}	claro que toda gente quer ver o real -mancheste r..porque o manchester joga a bola e quer ver o real a levar na pe i da
1984	{có co=cocó}	não será nada fácil para o 2º classificado nem para o có co rócococóoooooooooooooooooooooo!!!!
2004	{b0sta…=bosta}	isso só significa que o nosso campeonato é uma valente b0sta...
2337	{bô. Sta=bosta}	m' ta concôrda cu bô. sta ta poi culpa na árbitru ê disculpa di quênha qui ca sâbe pêrde cu classi. nu djuga más qui ês, mas nu tem qui sâbe ôdja 'ma ês ê más forti.
2349	{enrrabado!=enrraba}	cala-te enrrabado! reduz-te à tua insignificância!

Source: authors' own elaboration

Table 8. Confusion matrix for predictions on the modified Portuguese dataset

		ForumForte prediction		
		Profanity Positive	Profanity Negative	
Augmented labels	Profanity Positive	481(86%)	80(14%)	561
	Profanity Negative	45(2%)	1894(98%)	1939
		526	1974	2500

Source: authors' own elaboration

Conclusions

Profanity disguising in user–generated text exploits the robustness of the human mind to visually interpret the semantics of a message overwritten with substitutions of twin symbols or insertion of bogus segmentations. Lexicon–based exact comparison filters, on the contrary, have limited ability to detect such variants. Here we have briefly described a technique inspired on algorithms for sequence alignment widely used in bioinformatics, and its implementation as a software prototype to prevent this anomaly on a Web forum. Our empirical study of this software indicates its potential applicability as a tool for content moderation in different communities, both at vocable or sentence level. Its detection (classification) mechanism is language independent and requires no training before use other than setting up a lexicon of plain forbidden terms.

There remain several issues that need further investigation in future work. One particular aspect is related to refining the mechanism so as to improve its effectiveness in detecting guises of very short profanities (sequences with three or less symbols), as in both of the tested datasets in different languages, these cases proved hard to identify. Adapting the lexicon or tuning the tolerance parameter τ automatically by learning from changing disguising patterns adopted by the users is another interesting avenue of research (e.g., bagging filters with different tolerances or tuning the costs of different edit operations). Besides, deployment of the software on large-scale real-life content-generation environments, considering the associated algorithmic or computational issues for speed and concurrency, is also appealing.

Besides, it would be worth exploring further scenarios for potential application of the tool. Obvious choices are social networks such as Twitter, Facebook and Instagram. Offensive comments in these digital platforms may lead to more severe gender or psychological incidents of cyber-bullying, racism, or sexual harassment. Recent studies have shown that these are becoming prevalent and serious problems [2], [18], [19]. On the practical side in this line of work, we anticipate our tool may contribute as a content pre-processor, for example to reconstruct corrupted comments (because of intentional disguising or involuntary misspellings) that can be then used as input for other information extraction and machine learning techniques for content classification, such as those described in [20]. On the research side, such studies are now routinely relying on crowdsourced labelling of large scale datasets [18], [19], precisely because of the difficulty of single-handedly labelling such very high volumes of comments; our tool can help to validate, augment, or automatize the labor of human annotators so as to minimize the risk of overlooking positive cases, as it was shown in our study (Section 2.3).

Acknowledgements

We would like to acknowledge the anonymous reviewers for their constructive appraisal and suggestions for improvements on an early version of the manuscript.

References

[1] T. O'Reilly, "What is Web 2.0: Design patterns and business models for the next generation of software," *Commun Strateg.*, no. 1, p. 17, 2007.

[2] S. Sood, J. Antin, and E. Churchill, "Profanity use in online communities," in *Proc. SIGCHI Conf. Human Factors in Computing Systems, ACM*, 2012, pp. 1481–1490.

[3] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Cursing in English on Twitter," in *Proc. 17th ACM Conf. Comput. Supported Cooperative Work & Social Computing*, 2014.

[4] M.-E. Maurer and L. Höfer, "Sophisticated phishers make more spelling mistakes: using URL similarity against phishing," in *Cyberspace Safety and Security*. Berlin: Springer, 2012, pp. 414–426.

[5] S. A. Rojas-Galeano, "Revealing non-alphabetical guises of spam-trigger vocables," *DYNA*, vol. 80, pp. 15-24, 2013.

[6] X. Zhong, "Deobfuscation based on edit distance algorithm for spam filtering," in *Machine Learning and Cybernetics (ICMLC), 2014 International Conference on*, vol. 1. IEEE, 2014, pp. 109–114.

[7] V. P. Cardona-Zea and S. A. Rojas-Galeano, "Recognizing irregular answers in automatic assessment of fill-in-the-blank tests," in *Engineering Applications (WEA), 2012 Workshop on*, 2012, pp. 1–4.

[8] S. A. Rojas-Galeano, "Towards automatic recognition of irregular, short-open answers in Fill-in-the-blank tests," *Tecnura*, vol. 18, 2014.

[9] C. Mogollón Pinzón and S. Rojas-Galeano, "A genomic-based profanity-safe Web forum," *Proc. 10th Colombian Computing Conference, IEEExplore*, 2015.

[10] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.,* vol. 48, no. 3, pp. 443-453, 1970.

[11] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, 1981.

[12] D. Venema. "Evolution basics: Genomes as ancient texts". *The BioLogos Forum*. [Online]. Available: http://biologos.org/

[13] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21, pp. 168–173, 1974.

[14] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov Phys Doklady*, vol. 10, no. 8, 1966.

[15] A. Leff and J. T. Rayfield, "Web-application development using the model/view/controller design pattern," in *Enterprise Distributed Object Computing Conference, Proc. Fifth IEEE Int.*, 2001.

[16] D. Alur, D. Malks, J. Crupi, G. Booch, and M. Fowler, "Core J2EE Patterns (Core Design Series): Best Practices and Design Strategies". Sun Microsystems, 2003.

[17] G. Laboreiro and E. Oliveira, "What we can learn from looking at profanity," *Computational Processing of the Portuguese Language*. Berlin: Springer, 2014, pp. 108-113.

[18] P. Burnap and M.L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, no. 1, pp. 1-15, 2016.

[19] H. Hosseinmardi *et al.*, "Analyzing labeled cyberbullying incidents on the Instagram social network," *Social Informatics: 7th Int. Conf. (SocInfo 2015)*, Beijing, China, December 9-12, 2015, 2015, pp.49-66.

[20] S. H. Yadav and P. Manwatkar, "An Approach for offensive text detection and prevention in social networks," *Innovations in Information Embedded and Communication Systems (ICIIECS), 2015 IEEE 2nd International Conference on*. IEEExplore, 2015.