Flames, Risky Discussions, No Flames Recognition in Forums

Maria Teresa Pazienza[a], Armando Stellato[a], Alexandra Tudorache[ab]

a) AI Research Group, Dept. of Computer Science, Systems and Production
University of Rome, Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
{pazienza,stellato,tudorache}@info.uniroma2.it

b) Dept of Cybernetics, Statistics and Economic Informatics, Academy of Economic Studies
Bucharest
Calea Dorobanţilor 15-17, 010552,
Bucharest, Romania
alexandra.tudorache@gmail.com

Abstract

In this paper we describe our experimental study on flames and risky topic recognition. Firstly we introduce our approach, experiments and research goals. Then we provide basic definitions of flame, risky discussions and their difference. Furthermore we will analyze the most important features of flames by highlighting general, Italian and English features. Then we will focus on experiment and corpus description concluding by commenting the results of our test.

1. Introduction

World Wide Web supports the creation of virtual communities and interactive websites at a large extent. Forums and discussion boards represent a very important and significant social phenomenon inside the WWW. The traditional face to face discussions are migrating into the Web with new and limited rules for maintaining person to person interaction in a polite mainstream. Previously unknown persons discuss about several topics, while each participant contributes to the discussion with his own background, culture and language comprehension: totally uncertain contexts emerge!

Moreover, in those environments, traditional important elements of communication (such as gestures and voice tone) that underline emotion and sentiments are missing. Sometimes they are replaced by *emoticons* (icons that express emotion) but, for example, they do not wear voice inflections. As a consequence misunderstandings and flames occur in forums and discussion boards at an higher rate than in face to face conversations. Moderators are forced to block discussions between involved persons and interaction fails.

To maintain democracy on the web while avoiding unpolite and violent discussions, it is important for forum and discussion groups administrators to be able to recognize possible source of flames before they reach an undesirable state, without the need of constantly checking whole message boards.

We have analysed several forums in order to verify the existence of specific contexts enabling us to recognize the phenomenon. In the literature, at a border line between linguistics, psychology and cognitive science, we found a few contributions that have been revisited and enriched with classification algorithms in a computational approach.

Multidisciplinary studies have been considered to develop our approach and set up an experiment.

As we will use a specific terminology for which there not exists still a common understanding, by first we will provide definitions for specific terms object of our analysis. Then the adopted methodology, the test corpus and results will be widely discussed.

2. Flame, Forum, Risky Discussion, Topic and Post Definitions

A *forum* is a virtual "place" that hosts several discussions considered as written conversations. Forums can be either generic (where discussions span over many topics) or specific boards (where participants can talk about only one domain like politics, art, religion etc).

We define the *topic* as one discussion thread with one subject and more interactions (posts).

A *post* is the response provided by a specific user. Usually forums require login and it is not possible to post anonymously.

A *flame* is defined by a sequence of "non constructive" posts, with no positive contribution to the discussion. In flames users attack each other at a personal level instead of contrasting the dialog partner for his/her approach, contribution to the discussion or argumentation. As a consequence the flames phrasal structure is closer to oral than to written dialogue. Flames often induce moderators to close discussions.

We do a neat distinction between flames and risky discussions; in fact *risky discussions* contain a few flaming elements but could not be categorized as flames. They express a sort of "expectation" for a flame!

In real life the distinction between flames and risky discussions is usually subjective. It depends on moderator skills, his attitude, level of stress and level of implication in the subject. Sometimes the moderator himself generates flames due to his tough policy.

For our experiments we recognize a flame as a sequence of

six to eight posts that disrupt topic and represent personal attacks, while a risky discussion is considered as a sequence of three to four posts of personal attacks.

3. Flame and Risky Topics Features

We widely analyzed several forums either in English or in Italian languages in order to identify the existence of "flame features". By accessing also to a few psychological studies (King, A., 1995 ; Bucci, W., Maskit, B., 2005; Leahy, S., 2006) we have been able to highlight specific linguistic behaviors in flames. To start we collected language independent features, then language specific ones (Italian and English). It seems, in fact, that differences in cultures determine different ways for flaming.

General flame features:
1. Discussions take place between two or maximum three users on a specific topic or over more topics.
2. Short posts without much argumentation.
3. Posts made by new users (newbies) that are not able to integrate from the beginning into that specific community. (King, A., 1995)
4. Offensive language.
5. Phrase or expression ambiguity (eg. "un" against "il"; "bene", "niente" – Italian, "the" against "this", "the idea" against "my idea"- English); psychological clinical studies demonstrated that an ambiguous language is sign of tension. (Bucci, W. , Maskit, B., 2005)
6. Tough or nonlinear moderation policy.
7. Nonsense and off topic posts written by moderators in a no flame or risky discussion.
8. Off topic personal questions about other users of that forum; simple off topics usually doesn't determine a flame.
9. Non acceptance of community rules; it is related to few users that some times disrupt the activity of an entire forum. (Leahy, S., 2006)
10. Speed in posting; flames often take place almost in real time.
11. Repetitive cites from other posts. One user try to attack another user attacking every idea of that specific user.
12. Proper names are not relevant in a flame.

Italian flame features:
1. Users addresses directly to each other ("cut and thrust" - "botta e risposta" in Italian).
2. A few users attack each other frequently over more topics.
3. Use of short phrases, often without a subject (è una cosa stupida; non è vero - "it is a stupid idea; it is not true")
4. Hypocritical politeness in expressions like: "perdonami se...", "scusa se..." ("excuse me...")

English flame features:
1. Users attacking ideas instead of other users directly. (This is often not valid for sport/games and teens forums – see Notes below)
2. Users adopt a lot of ironic expressions instead of direct expressions.
3. Adoption of slang or urban expressions.

Notes: Language in sport/games and teens forums is usually biased. In these contexts, flames are recognized by user behaviors more than by language models.

Risky topics features:
We are interested also in recognizing possible risky situations, then we tried to find general contexts in which they occur. In the experimental set up, they are used as a third class in between flames and no-flames; in fact risky situations, while not being flames, reveal a few features overlapping with those of flames.
1. As in flames, arguments take place between two or maximum three users on one topic or more.
2. Risky discussions are characterized by mini-flames (two or three flame posts) followed by normal discussion.
3. In risky discussions we can find more than one mini-flame; moderators usually don't make hush interventions.
4. In risky discussions are present impersonal constructions against personal construction as in flames; when a risky topic turns personal flaming occurs.
5. In risky discussions often we will find a good number of off-topic posts.

4. Algorithms for Flames and Risky Topics Classification

In the previous section we introduced flame and risky topics features. In this section we will see some methods for identifying flames and risky discussions in forums. There are two main techniques for emotion and sentiment classification, roughly: symbolic and machine learning techniques. The symbolic approach may use manually defined rules, lexicons, conceptual knowledge,… where a machine learning approach uses unsupervised, weakly supervised or fully supervised learning to construct a model from a training corpus. (Basili, R., Moschitti, A., 2005)
Once agreed on flame feature classes (as naively summarized into previous sections) and completely structured them, we will set up a dedicated knowledge based architecture to recognize flames in forums. In fact, while we analyze discussions in forums as a textual sequences, posts contain a lot of typical oral/gergal expressions. There is a mixture of written oral structures and components thus requiring a dedicated processing.
Meanwhile we are verifying the possibility of using

machine learning algorithms as a first approach.

The main applications of supervised machine learning are classification algorithms. We will focus on document classification.

For document classification we can use classic supervised learning techniques (e.g. Support Vector Machines, Naive Bayes, Maximum Entropy).

An efficient yet simple classification algorithm is Naive Bayes; it does the naive assumption of document features independence. Later on this document we will present some experiments on two and three document category classifier.

Unlike in supervised learning, in case of unsupervised learning methods is not needed the manual labeling of inputs (the so called training set).

A well known unsupervised method is clustering that not always is probabilistic.

A method that allows to vary the number of clusters with the problem size and also gives to the user the control over each cluster member similarity degree is the adaptive resonance theory (ART).

ART networks are used for many pattern recognition tasks. The first version of ART, "ART1", was developed by Carpenter and Grossberg in 1988.(Manning, C. D.; Raghavan, P.; Schütze, H., 2008)

4.1. Our Approach

For our experiments we used the Bayesian Classifier with Laplace smoothing supervised approach.

Naïve Bayes algorithm is based on the assumption of events independence. Although it is an "incorrect" assumption that documents features are completely independent, the Naive Bayes algorithm proved over time to have a very good performance in text classification (Basili, R., Moschitti, A., 2005; Manning, C. D.; Raghavan, P.; Schütze, H., 2008; Yu Bei; Unsworth J., 2007). Moreover, many of our analyzed features (e.g. length of the post, newbie recognition etc...) relate to information which goes out of the textual boundaries of the documents and is thus more prone in being used with such assumption.

4.2. Experimental Setup

For our experiments we adopted the "Waikito Environment for Knowledge Analysis" Weka set (Witten, I. H.; Frank, E., 2005) for using a Naïve Bayes algorithm implementation and an extension of Word Vector Tool for producing word vectors by extracting data from Forum sources.

By default Naïve Bayes in Weka's implementation uses Laplace smoothing: always adds 1 to the number of different values for a particular attribute.

5. Experiment Description

In this experiment we will verify the possibility to classify flames *versus* noflames as well as risky discussions. In fact, forum administrators are interested to "prevent" (moderate risky discussions) rather than to "cure" (close flame topics).
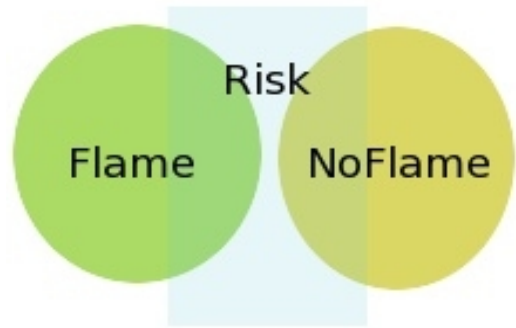

Figure 1: Topic Class Diagram

We designed two sets of experiments: one to recognize flames and another to recognize either flames or risky discussions.

Flames and noflames are two disjoint classes while (as discussed in section 3) risky topics class has features from both flame and noflame classes thus representing a sort of fuzzy class.

For training purposes risky topics were extracted from noflame category meanwhile flame category remains unchanged during both tests.

5.1. Corpora Structure

From observing several forums and discussion boards we learned that flames usually are generated by few users that express their ideas using a rather aggressive language. Moreover each forum is a community with its own rules, participants, language, interests, moderation and administration staff. Each of these elements impact significantly on flames characteristics.

For data homogeneity and evaluation purposes, we must use the same forum both for training and testing. The Italian corpus is represented by sequences of posts extracted by topics found on **postare.it**. We choose this forum as it is general, rich of politics and daily matters and has the right content structure for our experiments: normal/noflame discussions (65%), flames (5%) and risky discussions (30%).

In forums (as in any virtual or real community) the level of emotionality changes over time and so does the forum structure. Hereafter we provide a few examples from training corpus in Italian language.

All the examples have been translated into English preserving as much as possible of the original sense and style to better illustrate emotion. We tried to maintain phrase structure and grammatical non concordances, too.
 " by Gregoriosurace
Testo Quotato:Postato originariamente da Ilpiero :
Sfogo amaro condivisibile ma...
Testo Quotato:da GregorioSurace: Complimenti comunisti.

- mi spieghi cosa c'entra? Senza polemiche o altro... A una persona che mi chiede una delucidazione non posso fare altro che dargliela. Evidentemente il mio testo non era abbastanza chiaro. Anche se a me sembrava cosi..." -
"I understand your bitterness but...
Quoting Text: GregorioSurace: Compliments to Communists. - Could you explain? No polemics please... If someone asks me something, I must answer. Obviously my post wasn't clear enough. Although it seemed to me so ... "
(Excerpt from training **flame** 1779.txt)
"by Francesca

---

tutti noi abbiamo la nostra libertà di pensare ciò che vogliamo di chiunque.
ma c'e anche da dire che io non ti volevo convicere in nulla io ti stavo solamente dicendo che come tu non hai bisogno di niente, meglio per te tu hai il tuo stile di vita ma ognuno di noi hanno dei vizi chi il fumo, chi alcol, chi spendere i soldi, chi riempire il cu+o allo stato combattendo le loro battaglie e poi questo e il mondo in qui viviamo, un mondo di guerre combattute dai poveri e in tanto lo stato si fa i villoni e i viaggietti con l'aereo privato...ma lasciamo perdere....
la droga esiste da sempre...secondo te gli indiani nella pipa della pace cose c'era tabacco? anche nel fumo ci sono delle regole da rispettare e se propio lo voi sapere da come gira il mondo ora preferisco uno che si fa una canna al giorno che uno che crede un qualcosa che non e!

_____

by Asmodeus

E tientelo pure guarda, fai un favore a tante altre persone che invece non lo vogliono "
"By Francesca

We all are free to think what we want of anybody. But I also want to say that I didn't want to persuade you on anything. I was just saying that much the same way you don't need anything, it is much better for you, you have your own lifestyle but everyone have their vices like smoking, drinking, squandering money, fill the state *** fighting their battles and thus this is the world we live in, a world of wars fought by the poors and where the state builds villas and travels by private jets... but let's go over it ....
Drugs have always existed ...do you believe in the Indian peace pipe there was tobacco? There are rules to be respected even for smoking and if you'd like to know my opinion I prefer a person who smokes drugs once a day to one who believes in something that is not!

_____

 By Asmodeus----Look do us a favor and keep it"
(Excerpt from training **risky** discussion r_2151.txt)
As it is evident, there is a very narrow difference between flames and risky topics and this is reflected in the experiment results as we will see in results section.

5.2.Training and Testing Notes

The annotation of each sequence of topics was made manually. As stated in section 3 flames were extracted from integral topics as 6-8 flame post sequences, risky discussions as sequences of 6-8 post in which 2-3 post sequences were mini-flames.
Testing was made on the same forum. Corpus was preprocessed as for the training. After manual classification the training and testing documents were selected randomly to guarantee a fair test.

5.3. Results and Evaluation

For evaluation purposes, as in information retrieval, the two main indicators - precision and recall - have been used.

$$Recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$Precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

For our experiments we used a weighted precision and recall.
In forum administration is preferred to recognize all flames and have some false positives than to not completely recognize flames.
We considered "error weight" as an index of the gravity of the error. Flames recognized as noflames and noflames as flames errors are considered "full errors" and have error weight=1. Risky topics recognized as flames and risky topics recognized as noflames are considered "partial errors" and we attributed them an error weight of 0.5.
So for each experiment precision and recall formulas have been rearranged as in the following:

$$Recall = \frac{tp}{tp + \sum fn \times ew}$$

$$Precision = \frac{tp}{tp + \sum fp \times ew}$$

Where:
tp = true positives - topics correctly identified
fn = false negatives - correct topics that have not been found
fp = false positives – incorrect topics classifies as positives
ew = error weight

6. Experiment Results

In this section we analyze results of a couple of experiments in flame, no flame and risky topics identification.

6.1. First Experiment Results: Flames and No Flames Classifier

In the first experiment we were interested in flame and no flames classification.

The training set was composed by 35 flames and 95 noflames. The testing corpora was composed by 12 flames and 97 noflames including risky topics. Test results are shown in Table 1.

Flames identified as noflames	0
No flames identified as flames	18
Risky topics identified as flames	3
Flames correctly identified	12
No flames correctly identified	76
TOTAL FLAMES TESTED	12
TOTAL NO FLAMES TESTED	97
TOTAL TOPICS TESTED	109

Table 1: First experiment results: flames and no flames recognition

As you can see in Table 1 we identified all flames and a good number of noflames - 76 from a total of 97: about 78% .

In Figure 2 we presented the correctly identified topics by category.
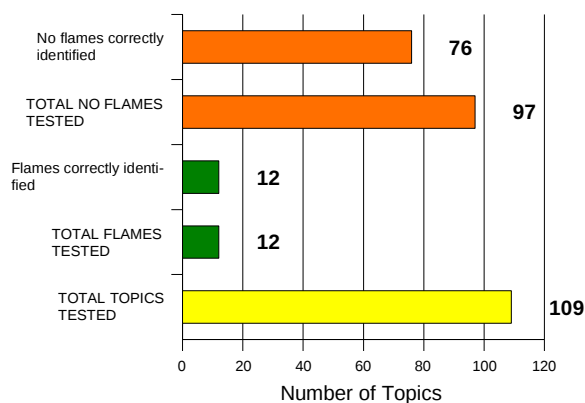

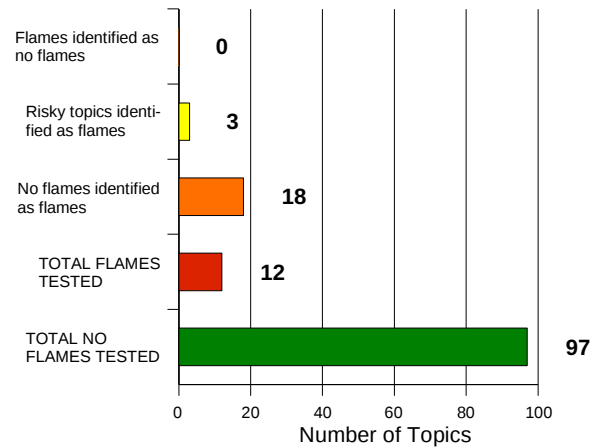Figure 2: First Test Correctly Identified Flames

Figure 3: Errors in Topic Classification

In Figure 3 you can see the errors of topic classification. In conclusions and future work section we will present some methods to minimize such errors.

While we obtained a complete recognition of flames, we get some false positive b:most of them could be classified as risky discussions by a manual inspection.

E.g.

" by Mime

....appena sentito al tg3.... il nanetto "non c'e stato alcun editto bulgaro, il mio era un appello ai nuovi dirigenti che stavano per insediarsi in rai che "certe cose" non si verificassero piu"

mi domando e dico esistesse un limite all' idiozia e alla sfrontatezza di quell' omuncolo?

criminoscio "

"By Mime

.... Just heard on tg3 .... The little one "there 'was no Bulgarian edict, mine was just an appeal to the new leaders coming into RAI... that" certain things "shouldn't happen any more"

I wonder if there is a limit to stupidity and impudence of that little man! Criminoscio" (Excerpt from __risky discussion r_1705.txt)__

The two main indicators are the following:

Recall	1
Precision	0.53

Table 2: First Experiment – Flame Recognition Precision and Recall

Precision and recall indicates that on this first classification we have the ability to find all flame topics and that the probability to find the most relevant topics first is over 50%. From a forum administrator it could be a good result: in fact forums are very dynamic contexts and used language is not

standard at all - each participant uses his own language, abbreviation ways to express disappointment etc.

Other studies have reached a lower recognition rate like Ellen's Spertus Smokey. (We have not been able to find further references on flame recognition). In Table 3 we reported the comparative results of our test based on Naïve Bayes Smoothed and her tests based on C 4.5 algorithm. (Spertus, E. , 1997)

	Naïve Bayes Smoothed	**C 4.5**
Flames Recognition (%)	100%	39%
No Flames Recognition (%)	78%	97%

Table 3: Comparative Results of Naive Bayes and C 4.5 Algorithms

6.2. Second Experiment Description: Flames, Risky Topics and No Flames Classifier

For the second experiment we annotated three classes: flames, noflames and risky discussions.

The training set was composed by 35 flames, 95 noflames and 22 risky discussions. The testing corpus was composed by 12 flames, 88 no flames and 9 risky topics.

Once again we get no false negatives on flames/noflames and only two false negatives on flames/risky topics identification as shown in Table 4.

FLAMES

Flames identified as noflames	0
Flames identified as risky topics	2
Flames correctly identified	**10**
No flames identified as flames	19
Risky topics identified as flames	3
No flames+risky topics correctly identified	*75*
TOTAL FLAMES TESTED	**12**

Table 4: Flame Topics Identification

After a manual inspection, the two flames' false negatives classified as risky discussions revealed to be not very aggressive ones. It is remarkable that no flame was identified as noflame.

The risky topics identification is the weakest test: it is not a surprise due to the nature of risky topics. Risky topics have both elements of flames and of noflames topics and by definition it is not a disjoint class as shown in Figure 1.

For risky discussions identification we have only one true positive from 9 tested, 3 topics were identified as flames and 5 as noflames (see Table 5).

RISKY TOPICS

Risky topics identified as flames	3
Risky topics identified as no flames	5
Risky topics correctly identified	**1**
Flames identified as risky topics	2
No flames identified as risky topics	2
No flames+flames correctly identified	*77*
TOTAL RISKY TOPICS TESTED	**9**

Table 5: Risky Topics Identification

No flames are correctly classified in a proportion of 76%. On noflames we have 21 false negatives of which 2 topics are identified as risky discussions.

NO FLAMES

No flames identified as flames	19
No flames identified as risky topics	2
No flames correctly identified	**67**
Flames identified as noflames	0
Risky topics identified as noflames	5
Flames+risky topics correctly identified	*11*
TOTAL NO FLAMES TESTED	**88**

Table 6: No Flame Topics Identification

We had 5 false positives but only risky topics identified as noflames.

In Tables 7,8 and 9 we show precision and recall for every class.

Flames Recall	91%
Fames Precision	33%

Table 7: Flames Classification Indicators

Risky Topic Recall	20%
Risky Topic Precision	33%

Table 8: Risky Topics Classification Indicators

No Flames Recall	77%
No Flames Precision	96%

Table 9: No Flames Classification Indicators

Flames Identification Recall raises to 91%. Risky topic identification instead has a poor precision and recall as expected.
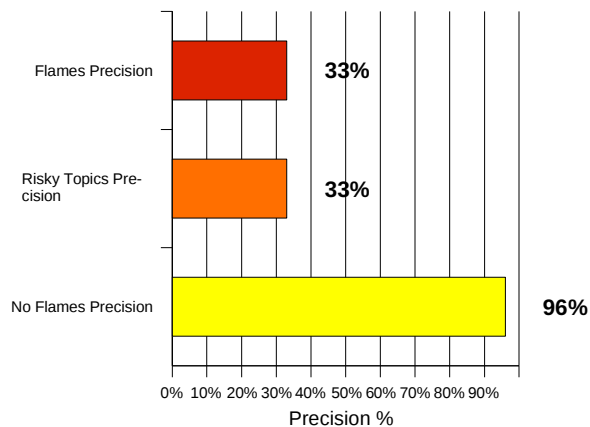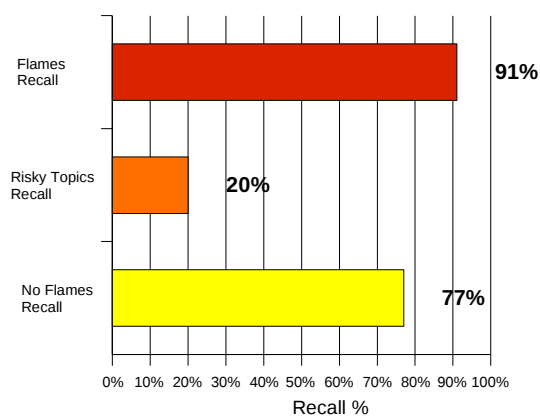
Figure 4: Second Test Comparative Precision

Figure 5: Second Test Comparative Recall

In Figures 4 and 5 we presented the comparative results of precision and recall for the second experiment.

Results clearly show the difficulty of risky discussions categorization and the duality of risky topics. Each topic has both elements of normal – no flame and flame topics.

Since risky topics present heavily dependent features their identification goes against Bayes' assumption of independence. So we need other methods to optimize results. A possible approach will be presented in Conclusion and future Work section.

7. Conclusions and Future Work

In this paper we presented our experimental study on flames and risky topic recognition later on definitions of flame, risky discussions and differences between flames and risky discussions.

Then we provided an exemplification of the most important features of flames and risky discussions distinguishing between general, Italian and English ones.

As widely described, we have got very good results in flames recognition and promising ones concerning risky

discussions identification. At the moment we are at an early stage in our research: we aim in searching further algorithms in order to better recognize risky situations. In fact forum administrators could benefit a lot by an early alert! The matter is very hard to formalize! We analyze discussions in forums as a textual sequences, while posts contain a lot of typical oral/gergal expressions. There is a mixture of written oral structures and components requiring a dedicated processing.

From our empirical observation we can verify that the easiest way to eliminate false positives is to   identify impersonal and personal phrase structure.

Impersonal constructions identify risky discussions whereas personal phrase constructions identify flames. Impersonal construction like "it is know", "is said", "si dice", "non e detto" indicates tension but not a direct attack to another user.

E.g. "tutti noi abbiamo la nostra libertà di pensare ciò che vogliamo di chiunque." - "we all are free to think what we want of whom we want" (Excerpt from __risky__ discussion n. 2151)

"non per stare sempre a parlare di lui ma il gesto del lancio di uova sembra l'inizio di una protesta che covava da tempo o un gesto isolato" - "I don't like to talk only about him, but the act of trowing eggs seems the beginning of a long underlaying protest or an isolated act".(Excerpt from __risky__ discussion n. 1803).

It emerges that in both sentence fragments predominate impersonal constructions while tension is sensible. These two are considered risky topics because they can degenerate easily in flames. Flames are characterized by personal constructions.

E.g. *"ma tu, hai una connessione wireless e per portare il tuo livello produttivo ai valori che cosi bene conosciamo..."* *- "but you, have a wireless connection and to bring your productivity to the levels we all know... " (Excerpt from __flame__ discussion n. 1651).*

"ah intendi dire che con quel post ... disturbavo qualcuno??"-"you want to say that with that post I was annoying someone?" (Excerpt from flame discussion n. 1826).

The heuristics about phrase constructions could be effective in distinguishing between flames and risky discussions.

In the future we plan study the possibility to integrate such heuristics in topic classification.

8. References

King, A. (1995) *Effects of Mood States on Social Judgments in Cyberspace: Self Focused Sad People as the Source of Flame Wars*, Storm, July 2, 1995, http://psychcentral.com/storm1.htm

Bucci, W.; Maskit, B. (2005, 1). *A weighted dictionary for Referential Activity. Computing Attitude and Affect in Text;*, 49-60

Leahy, S. (2006) *The Secret Cause of Flame Wars*, Feb 13 , http://www.wired.com/science/discoveries/news/2006/02/70179

Basili, R.; Moschitti, A. (2005) *Automatic Text Categorization: From Information Retrieval to Support Vector Learning*, Aracne Editrice, Informatica, ISBN: 88-548-0292-1

Manning, C. D.; Raghavan, P.; Schütze, H. (2008) *Introduction to Information Retrieval*, Cambridge University Press (to appear in 2008), http://informationretrieval.org

Boiy, E.; Hens, P.; Deschacht, K.; Moens, M. F. (2007) *Automatic Sentiment Analysis in On-line Text*, Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007

Dave, K.; Lawrence, S.; Pennock, D. M. (2003) *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In Proceedings of WWW-03, 12th International Conference on the World Wide Web, ACM Press, Budapest, HU, 2003, pp. 519–528.

Yu Bei; Unsworth J. (2007) *An Evaluation of Text Classification Methods for Literary Study*, University of Illinois at Urbana-Champaign, http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=157

Janssen, Jerom F., (2007) *Diachronical Text Classification, A study of text properties and their changes over time*, PhD. Thesis, University of Groningen, May 14.

Weka, Weka 3: *Data Mining Software in Java*, http://www.cs.waikato.ac.nz/ml/weka/

Witten, I. H.; Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition). Morgan Kaufmann

Word Vector Tools, *The Word & Web Vector Tool*, http://nemoz.org/joomla/content/view/43/83/lang,en/

Spertus, E. (1997) *Smokey: Automatic recognition of hostile messages*. In Proceedings of the Ninth Conference on Innovative Application o f Artificial Intelligence (IAAI-97), Providence, RI, pages 1058-1065. AAAI Press/The MIT Press, July 1997