A holistic system for troll detection on Twitter

A Holistic System for Troll Detection on Twitter

Abstract

Various techniques based on artificial intelligence have been proposed for the automatic detection of online anti-social behaviors, both in existing systems and in the scientific literature. In this article, we describe TrollPacifier, a holistic system for troll detection, which analyses many different features of trolls and legitimate users on the popular Twitter platform. In this system, the most known and promising approaches and research lines are applied, along with original new ideas, in a form that fits such a large public platform. In particular, we have identified six groups of features, based respectively on the analysis of writing style, sentiment, behaviors, social interactions, linked media, and publication time. As its main scientific contributions, this work provides: (*i*) an up-to-date analysis of the state of the art for the problem of troll detection; (*ii*) the systematic collection and grouping of features, on Twitter; (*iii*) the description of a working holistic system for troll detection, with a very high accuracy (95.5%); and (*iv*) a comparison among the different features, with a machine learning approach. Our results demonstrate that automatic classification can be useful in the whole process of identification and management of online anti-social behaviors. However, a multi-faceted approach is required, in order to obtain an adequate accuracy.

Keywords: Troll Detection, Machine Learning, Online Social Networks

1. Introduction

For a long time, it has been difficult to find an accurate description for an Internet "troll", because the act of trolling is strongly subjective. In part it still does not have a single definition, resulting in a poor comprehension and in a low

interest by the researchers' community. As a general understanding, a troll is an individual adopting an antisocial behavior that provokes and irritates other users of an online social platform, often using an aggressive or offensive language, and whose effect is to derail the normal evolution of an online discussion and possibly to stop it. Thus, "trolling" refers to "a specific type of malicious online behavior, intended to disrupt interactions, aggravate interactional partners and lure them into fruitless argumentation" [1].

Quite recently, this problem has attracted the public attention. In fact, renowned press agencies and magazines have begun to address the issue and to write articles both on the general description of the phenomenon and on particular events that have caused a stir about "toxicity" of some social media sites [2]. The need of dealing with this problem has emerged over time, along with some scientific studies conducted by various universities and with practical functionality provided by online platforms like Twitter, which periodically releases new services or new features, such the ability to report anti-social behaviors or to "mute" annoying users. However, all attempts tried until now have not been able to eradicate the problem. Hence, it is fundamental to create an automated method that leverages on artificial intelligence, data mining and social network analysis in order to manage the complexity of the issue. In particular, the main contributions of this work are:

- Up-to-date analysis of the state of the art for the problem of troll detection.
- Systematic collection and grouping of features, on Twitter.
- Construction of a working holistic system for troll detection, with a very high accuracy (95.5%).
- Comparison among the different features, with a machine learning approach.

The article begins with an overview on the state of the art on existing methods, cataloging them according to their approach. Secondly it illustrates an

adaptation of the best techniques identified within the context of Twitter, to review the most useful metrics for representing information from social media. These metrics, which are used as features for an automatic classifier, are compared and evaluated according to their relative importance for the classification process. Finally, the article ends with the final analysis of the results and the possible future developments in this field.

2. Related Works

2.1. Analysis of the online trolling phenomenon

With the rise of virtual communities, users of online services have been able to benefit from a new simple and fast way of communicating, suitable for connecting remotely separated individuals. This kind of Computer-Mediated Communication (CMC) can also provide varying degrees of anonymity that can encourage a sense of impunity and "freedom" from responsibility for users. This whole scenario has led to the development of a widespread phenomenon that occurs within the CMCs, known as *trolling*. The first references to the use of the word *troll* on the World Wide Web have been found in Usenet, a forum community popular in the eighties. A troll is generally defined as an individual who is marked by a negative online behavior [3, 4], or as a user who initially pretends to be a legitimate participant, but later attempts to disrupt the community, not necessarily in a blatant way, but with the effect of attracting the maximum number of responses [5]. Trolls are also described as individuals who derive pleasure from annoying others [6], and, in fact, recent researches have discovered that sadism is closely associated with those who have trolling tendencies [7]. In [8], the connection between dark personality traits and engagement in harmful online behaviors is investigated.

A troll seeks to cheat a person or a whole community [9]. In sociology, the term has become a synonymous for all negative online behaviors, but it is necessary to recognize each one by giving them a definition in order to understand and face the online trolling phenomenon in a systematic way. Studying the

behavior of some users within the virtual communities of Usenet, Hardaker [3] has found that the act of trolling is manifested through four interrelated ways:

- *Deception*: a troll will try to disrupt the group, trying to stay undercover; for example, when a troll intentionally disseminates false advices [5].
- *Aggression*: a troll that is searching for a conflict, can use a provocative tone towards other users.
- *Disrupt*: it is the act of causing a degradation of the conversation without necessarily attacking a specific individual.
- *Success*: often a troll is acclaimed by users for his degree of success, so trolling, despite being a nuisance for users, may end up at the center of attention of the group.

It is clear that trolling is a more complex problem than just provocative attacks. Although the concept may seem tied to the meaning of some words like rudeness, arrogance, impertinence and vulgarity, they do not provide an accurate description since typically trolling consists in keeping hidden the real intent of causing problems. In [10], the characteristics of troublemakers in online social networks are investigated. In addition, a recent study states that anyone can become a troll: in fact, their predictive model of trolling behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling [11].

These practices are often tolerated, in line with a common attitude on the Internet that considers insulting speech as a manifestation of freedom of expression [12]. In less vulnerable communities, with more experienced or emotionally detached users, some episodes can be also seen as playful actions. However, inexperienced or vulnerable users of online communities may feel trolling particularly painful, distressing and inexplicable.

To counter these actions, some services implement identity verification processes [13]. Nevertheless, the propensity to trolling seems to have become more

widespread recently [14]. A case study analysis of the behaviors and strategies of a group of alleged Twitter trolls is presented in [15]. In extreme cases, anti-social online behavior has also led to suicides of adolescents [16], thus, it is not surprising that this rising phenomenon is alarming the social network operators [17].

Even when trolling does not come as a direct attack, it can be a threat because it can manifest itself with subtler ways, for example as a mean to try to manipulate others' opinions. In fact, the rise of the Internet has allowed corporations and governments to freely disseminate false rumors, or to use other dishonest practices to polarize opinions [18]: it has been shown that a user's opinion can be influenced by other users' comments [19, 20].

Considering its diverse motives and forms, trolling represents a vexing problem in CMC, because it hinders the normal course of a conversation. Indeed, user contributions in the form of posts, comments and votes are essential to the success of an online community. But, with such a high degree of desired participation, excluding individuals with unpleasant online behaviors, as trolls, can lead to a perception of excessive control and censorship, trigger side effects, impede effective community development. Thus, the goal to protect discussion threads from trolling has to be accurately balanced with a certain level of tolerance, for avoiding unnecessary interruptions and facilitating the integration of novice and uneducated users.

2.2. Detection methods

Usually, online social networks rely on moderators for banning malicious users. In many cases, also common users are provided with the option to flag inappropriate posts and mute users. However, this kind of manual solution has some major drawbacks, including a delay of actions, subjectivity of judgment and scalability [21]. Thus, it is necessary to augment the process through some automatic mechanisms.

In Section 3, we will describe a troll detection system, composed by multiple automatic classification systems based on machine learning. However, to create

such a complex system, it is necessary to take into account the most distinguishing features of online trolls. For this purpose, this section systematically analyzes the relevant scientific literature. Taking into account this survey and literature review, Section 3 will provide a selection and a classification of the most promising features, to be used for creating a troll detection system.

Few research works consider some different aspects of online trolls. In a study on anti-social behavior in large online discussion communities (CNN.com, Breitbart.com and IGN.com), some general tendencies have been observed [22]. The analysis focuses on the users subsequently banned by the moderators, defined as "Future-Banned Users" (FBUs), and confronts them with more civil users, defined as "Never-Banned Users" (NBUs). Analyzing their behaviour before being banned, FBUs show a tendency to write comments which are difficult to understand, often off-topic and with an adversarial language [23]. They tend to focus on few discussion threads, but they contribute with more posts per thread and they also receive more answers than average, suggesting that success in attracting attention can be synonymous with abnormal behaviors. Furthermore, FBUs have high rate of post cancellation (by moderators) and signaling (by other users), increasing over time. The described system is able to predict when an individual will be banned, with over 80% of accuracy, analyzing four sets of characteristics: post content, user activity, reactions of the community, moderator's actions. A similar study [24] has been conducted on the community of an online newspaper (Dnevnik.bg). Authors have derived specific metrics, including: community rating, consistency with the topic, order of comments, answers, time of the day.

A definition of troll as somebody who was called such by other people was used in [25] to predict, in news community forums, whether a comment is written by a troll or not. In this work, most of the features are based on textual attributes, but they are evaluated in a good methodical way. In fact, the authors report the results of classifiers trained (i) using all features, as well as (ii) excluding one individual feature group. We have performed our set of experiments in a similar way, but considering more features, and in more varied

groups (Section 4). This way, it is possible to analyze distinct groups of features, related to the different aspects of troll definition and behavior that exist in the scientific literature.

The majority of research works in this area focus their analysis on few homogeneous features. To study this large variety of proposed analyses and research works systematically, we have identified six main types of approaches. For each type of approach, we have defined a group of features, using ideas proposed in previous researches as well as new ones.

This systematic survey of the scientific literature is not intended as a mere study or a reasoned comparison, but instead it is intended as the first step for creating an online automatic troll detection system. In fact, the approaches and features described in this section constitute the basis for the system, which will be described in the next section.

Sentiment analysis as well as other forms of text analysis work on a single content element. However, our aim is to classify troll users, instead of individual tweets and comments. Thus, we have aggregated results for all the texts in a user's timeline. In fact, our system includes multiple levels of analysis. Each method of text analysis that we have considered is applied to individual tweets and comments; then its results are aggregated to obtain various features that constitute the input of the subsequent analysis, which takes into account all behaviors and features of a given user.

2.3. Sentiment analysis

Some research studies apply sentiment analysis to the problem of troll detection. For example, in [26] sentiment analysis is applied on the Twitter social network and it is used to identify political activists hostile to other parties and to evaluate the degree of conflict between two different factions, during the 2013 electoral period in Pakistan. The researchers use a tool called SentiStrenght, which estimates the "force" of a sentiment (either positive or negative). In [27], another study is reported, likewise characterized by the analysis of political discussions on Twitter, which tries to spot the malevolent users through the con-

tent of their tweets. Using a similar approach, the VaderSentiment library [28] is based on a lexicon sensitive to both the polarity and the intensity of sentiments of words. It has been validated by multiple independent human judges and is tailored especially for microblog-like contexts; nevertheless, according to its authors, it is also applicable in other domains. Another proposed paradigm for text analysis is "sentic computing" [14]. This paradigm is more focused on semantics rather that syntax and it is more inclined to evaluate the sense of the text, including what is expressed implicitly. In fact, this model is not shaped on static learning models, but it uses tools based on domain specific ontologies.

In [31], an emotion detection system is described. The system is based on a hierarchy of classifiers, at three levels. The classifiers at the three levels distinguish, in order: objective / subjective tweets; positive / negative tweets (among the subjective ones); tweets expressing fear / anger / sadness (among negative tweets), or love / joy / surprise (among positive tweets).

The work illustrated in [29] tries to estimate, solely with metadata, the presence of trolls inside the reddit.com portal, and highlights some characteristics according the criteria set out above. All the obtained information is collected in attributes of instance variables used to train a Support Vector Machine classifier. Once tested, it has shown a good accuracy of about 70%. The results show that the approach based exclusively on metadata is less accurate than the ones based on the sentiment analysis but a combination of the two could bring benefit to both methods, like, for example, it happens in [30].

2.4. Time and frequency of actions

Frequency of publication has been related to the quality of online discussions by various studies. In [22], the features of users later banned from some large websites are studied. In addition to the kind of produced text, also patterns of activities are observed. It is found that useful features, to distinguish future banned users, include the frequency of some activities, as the number of posts and comments per day. In [32], newsroom interviews, reader surveys and moderators' choices are used to characterize the comments published on a newspaper

website. It is found that frequency of commenting is a valuable indicator of low quality discourse.

In [25], authors describe two classifiers: one for detecting "paid trolls", who try to manipulate a user's opinion, and one for detecting classical "mentioned trolls", who offend users and provoke anger. Among many features regarding sentiment and text analysis, based on lexicons and bag of words models, they also consider some metadata, including the publication time. In particular, they distinguish a worktime period (9:00-19:00h) and a nighttime period (21:00-6:00h). They also distinguish workdays (Monday-Friday) and week-end days (Saturday and Sunday). In fact, this kind of feature is found to have the largest impact on accuracy, according to this study.

2.5. Text content and style

Various approaches have been studied to carry out troll detection through the evaluation of the textual content of online messages. Some studies are based on the evaluation of the ARI (Automated Readability Index) of published texts, since it has been shown that a troll is more likely to write in a less comprehensible language compared to a normal user [22]. According to [29], a troll is more likely to write short comments, maybe because he writes faster replies compared to a non-malevolent user that writes more elaborated and longer sentences.

Other studies attempt to bring the troll identification problem to a higher level of analysis, studying not only individual messages, but entire discussion topics. This hybrid approach incorporates some of the techniques described in the previous subsection, but also adds new information obtained from the context in which the messages are integrated. Among them, [33] adopts a combination of metrics of a statistical and syntactic nature, and other elements related to the users' opinion: some of these measurements are similar to the ones already treated. Others manage to summarize more general properties of the discussion, like the number of references to other comments, how many times a determinate post is mentioned in the topic and the degree of similarity between the terms involved in the thread, which is a measure used also in other

studies and obtained thanks to the cosine similarity [22, 24]. The approach conceived by [23] is made by evaluating the problem from the same point of view, but using different concepts. It is based on the Dempser-Shafer theory [34], a generalization of the Bayes' probability concept, that turns out to be a very useful tool when it comes to imprecise and uncertain information, like the ones provided by the users of these environments. The study underlines how it is possible to characterize messages according to their apparent rationality, their degree of controversy and their relevance for the topic of discussion.

2.6. User behaviors

The necessity for integration of user level metrics for the problem of troll detection has emerged from various research works, e.g., [27]. In [22], authors focus their efforts on the extraction of users' general data. The aim is to study the most significant parameters for the characterization of a troll, thus obtaining a better perspective of troll behavior. In [35], various metrics are described, to measure a user's involvement in the platform and the nature of his/her participation. Some of the described metrics aim at distinguishing active users from passive ones, by comparing the number of original tweets and replies produced, with the number of retweets, quotes and likes.

In [36, 37], the different problems emerging from the interactions of users with online bots are tackled. In [36], an approach inspired by the biological DNA is applied to the analysis of users' behavior on social networks. In this case, sequences are constituted by codes representing different types of social actions, namely comments, likes, shares, and mentions. While the particular behaviors of bots and trolls may largely differ, both aim at diverting attention from the discussion topic. Thus, detection methods developed for one kind of abusive behavior may also prove useful for the other one.

2.7. Community level

This approach tries to solve the problem of troll detection through the study of the relationships within the online community, using the methodologies of social network analysis. To our knowledge, the first study which explores this field

is reported in [38]. In the study, troll detection is just a part of a comprehensive analysis of Slashdot Zoo, a portal that allows each user to label others as friends or foes. Thanks to this peculiarity, the social graph has some links with negative weights, which represent distrust and are useful to identify unpopular users. For troll detection, the most useful metrics are obtained through a variation of the Page Rank algorithm, taking into account negative weights, and by the raw number of foes of a node.

In [35], a modified version of the Hirsch Index is proposed for measuring the influence of a user. The Hirsch Index (h-index) is used in the research community to evaluate the scientific production of a scholar, on the basis of the received citations. In the context of Twitter, it can be defined as the largest number n, such that n tweets of a user have been retweeted or liked at least n times.

In [39], authors explain how to transform any social network in one with "friends and enemies". As a result, several solutions to troll detection based on this approach were born. For example, in [39] researchers try to improve this method by implementing an algorithm that, at each iteration, reduces the size of the social network by eliminating all edges that are unnecessary for the analysis, and focusing more on the types of "attacks" adopted by trolls. Instead, the work shown in [21] evaluates how it is possible to use the propagation of trust and distrust for measuring the reliability of a node.

2.8. Advertisements

Especially in the case of propaganda agents and opinion-spreading trolls, links to external content, like images, videos and articles, pay an important role [40]. This can be the case of paid trolls, political activists, influencers and advertisers. Advertisements of this kind include links to external content, but also to groups, pages and hashtags, often used to identify and mount viral campaigns. In fact, in recent years, social media are increasingly being used for creating coordinated and multi-faceted campaigns [41, 42]. Those activities include the role of human influencers, both willing and paid, troll users who

try to disrupt the discourse of adversaries or to attack opponents personally, bots, external content creators and news outlets. Thus, the presence of many forms of content advertisement can be taken into consideration for detecting and managing anti-social behaviors, in general.

3. TrollPacifier

On the basis of the works analyzed in the previous section, we can say that a user has to be considered a troll if his/her activities are driven by an anti-social behavior. Therefore, an ideal approach for identifying such a kind of users makes use of data at the user level. But our evaluation also tries to extrapolate additional features from the other approaches. In fact, this is suggested in most of the reviewed works. Nevertheless, no studies in the literature have tried to comprehensively explore this road. So, another goal of this work is to assess the compatibility of the various methods, integrating user-level metrics with features derived from the analysis of published texts and local social graphs. In this section, we will describe the process of data acquisition and the specific features used for the automatic classification process, adapted to the particular context of the Twitter social network [1].

3.1. Actor-based System

For realizing the TrollPacifier system, we have used ActoDES, which is a software framework which adopts the actor model for simplifying the development of complex distributed systems [43]. Actors are autonomous and concurrent objects, each characterized by a state and a behavior and the ability to interact with other agents through the exchange of asynchronous messages. After the analysis of its incoming messages, an actor can send more messages to itself or to others, create new actors, update its state, change its behaviors, terminate its own execution, etc. Each behavior can define a policy for handling

[1] `http://twitter.com/`

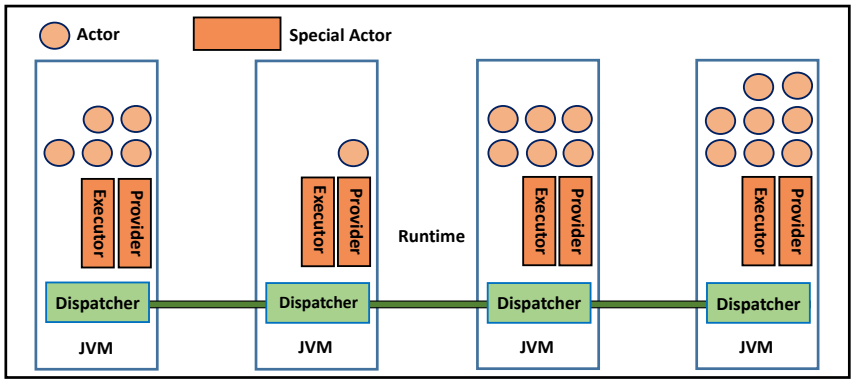

Figure 1: Distributed ActoDeS application architecture.

incoming messages, through handlers called "cases". Each case can only process messages corresponding to a specific pattern. An actor space is intended as a container offering the services needed for the correct execution of a set of actors. In particular, it includes two types of actors: a scheduler and a service provider. The former has the duty to handle the concurrent execution of actors, while the latter provides runtime services, needed by actors to complete their tasks. A subscription service is also available, to facilitate the development of collaborative applications with actors, as shown in [44]. This service has the task to receive incoming messages into a specific mailbox, and forward them to subscriber actors. The actors can eventually handle the messages differently, according to their own behaviors.

Other services, developed for this project, provide additional functionalities to actors, for the continuous analysis of various social streams. In particular, a Twitter service allows other actors to send various kinds of requests:

- User timeline, to obtain the recent tweets of a specified user and save them in a local storage system.
- Content query, to similarly obtain recent tweets published on Twitter and selected according to some constraints specified by the actor.
- Stream, to continuously receive messages published on the platform during

the execution; tweets are obtained in the form of JSON objects, which are then storea in a NoSQL repository (namely a MongoDB database [45]).

Leveraging ActoDES and the additional mentioned services, we have built a software system which can be used to track and study a news feed from social media, with an architecture that can be extended to different cases and also to more complex problems.

3.2. Data acquisition

The creation of a dataset of troll users is a crucial point of the analysis. In order to collect our training set we have used two cascaded approaches. The first one is based on distant supervision [46, 47] and allows one to obtain a raw dataset. The second one consists in manually filtering the previous dataset in order to obtain a more accurate training set.

In the first approach, we adapt an idea described by Mihaylov et al. [24] that defines a troll as a user that is called in this way at least N times by N different users. Twitter provides a series of official accounts, to which members can report their problems (e.g., @Twitter, @Support, @Safety, @TwitterUK, @TwitterAU, @TwitterSA, etc.). In particular, whenever a common user feels annoyed or even threatened by another, he/she can report the incident to one of these accounts, via a tweet containing a mention to the harasser, in the hope that Twitter administrators take the necessary measures. However, moderators are not always able to take appropriate countermeasures and often many users continue their online activities without being removed.

Therefore, we use the Twitter "Advanced Search" function to select the users which have been reported by other users that accuse them to be trolls (in messages containing words such as "troll", "ban", "harass", "block", "stalk", or some common derivatives). In this way we build a raw dataset of trolls composed by users mentioned in these messages, but not yet banned by administrators.

For the non troll class, we use the same approach and we select users starting from general tweets containing common words such as "a", "an", "the", "and".

We finally obtained a dataset composed by 3000 troll users and 3000 non trolls. By manual inspection of a hundred of instances of this training set, we have found many errors. In fact, our estimation is that more than a quarter of users mentioned to the support channels do not behave in an anti-social manner, according to the tracts discussed in section 2.1.

Therefore, we have decided to manually filter this raw dataset in order to obtain a more accurate training set, composed by 500 troll and 500 non-troll users. This final dataset has been validated by multiple independent human judges, through the manual inspection of users reported to the official support channels. In particular, users have been selected after inspecting both their recent timelines and their role and attitude in prolonged discussions, where they were repeatedly mentioned as trolls.

3.3. Groups of features

In section 2.2 we have discussed a large variety of proposed analyses and research works, systematically. Consequently, we have identified six main types of approaches. For each type of approach, we have defined a group of features, using ideas proposed in previous researches as well as new ones. In particular, for building TrollPacifier we have identified 6 groups of features, which are listed in the following paragraphs.

- **Sentiment analysis (*SENT*).** This group includes **26 features**, to distinguish positive, negative and objective posts, but also to associate them with more precise emotions. About "sentic computing" [14], TrollPacifier includes the main results of the SenticNet library [48]: sensitivity; polarity; trollness; attention; pleasantness; attitude. Moreover, it takes into account the results of lexicon and rule-based sentiment analysis, using the VaderSentiment library [28]. In particular, from this analysis TrollPacifier gathers values representing the maximum, minimum and average levels of positive, negative and neutral sentiments; polarity; trollness. Additionally, TrollPacifier includes a whole hierarchical emotion detection system,

as described in [31]. In particular, the output of each level of classification is used to obtain a feature for the user-level analysis. Thus, collected features are: number of objective and subjective tweets; number of positive and negative tweets; number of tweets expressing one of the six basic emotions of Parrot's model [49]: fear, anger, sadness, love, joy, and surprise. Finally, TrollPacifier includes an ad-hoc text-based classifier for evaluating the overall abusiveness of a text, i.e., provoking others to finally report the author as a troll. The classifier is trained with two classes of texts: those written by alleged trolls and those written by normal users. More details of this classifier are provided in the next section.

- **Time and frequency of actions (*TIME*).** This group includes **57 features**, to identify the most active day hours and the time dedicated to each post. Considering the results presented in [22, 32, 25], after an optimization process, we have included in TrollPacifier features for representing the activity in daily intervals of 4 hours. We have chosen this time length after a thorough comparison, in which we have trained automatic classifiers based on different algorithms (K-nearest neighbors, Naive Bayes, Sequential Minimum Optimization, C4.5) [50] using different interval length. Features measuring the activities in intervals of four hours provided consistently the best classification results. In TrollPacifier, the time intervals are distinguished by single day (from sunday to saturday), and also grouped together for generic workdays and weekends. In addition to these metrics, additional features consider the frequency of actions in the recent timeline and during the whole user's presence on the platform.

- **Text content and style (*TEXT*).** This group includes **31 features**, to measure the grammatical correctness and the kind of language used in posts. TrollPacifier includes some features for taking into account the readability grades, based on various metrics [33, 22, 24, 23]: Kincaid, ARI, Colemaniau, FleschReadingEase, GunningFogIndex, LIX, SMOGIndex, and RIX. TrollPacifier also includes the following other features in this

class: average word length and sentence length, by number of characters, syllables, or words; number of long words and complex words; number of verbs in general and some auxiliary verbs in particular; number of conjunctions, pronouns, prepositions, articles, subordinations, either in the middle or at the beginning of a phrase; number of hapaxes and rare words.

- **User behaviors (*BEHA*).** This group includes **38 features**, to distinguish users participating more actively, i.e., contributing with original messages and media objects. Taking into consideration the experiences of relevant research works [27, 22, 35, 36], we have introduced into TrollPacifier a number of features to characterize a user's online behavior, including: total number of tweets, retweets, replies, favorites, citations and quotes in the timeline; proportion between active actions (original tweets and replies) and passive actions (retweets and quotes); count of various actions associated with a single item (e.g., number of replies to a single tweet); maximum repetitions of a single action.

- **Interactions with the community (*COMM*).** This group includes **34 features**, to highlight a user's integration within his group of followers and followees. To represent a user's relationships within his own community, TrollPacifier includes the following features [38, 35, 39, 21]: number of followers and followees; ratio of these numbers; ratio of tweets per follower; number of posts retweeted or favorited by other users; counts of given and received mentions; number of different users mentioned; counts of different actions, including retweets, replies, mentions, related to a single user or to a single tweet; h-index based on retweets, likes, and their sum.

- **Advertisement of external content (*ADVE*).** This group includes **38 features**, to count the number of references to diverse external content and other channels of discussion. To evaluate the possible usefulness of external links and other forms of advertisement for troll detection [40, 41, 42], we have added the following features in TrollPacifier: number

and frequency of urls in posts and comments, as well as in the profile information provided to the platform; number and frequency of published or advertised videos, images and other media; number and frequency of hashtags.

3.4. General feature extraction

Apart from the system-level ActoDES actors, TrollPacifier includes additional actors, as shown in Figure 2. They are dedicated to (*i*) basic tasks, like acquiring streaming data and users' profile information from Twitter; (*ii*) direct feature extraction tasks, with different actors for the six different groups of features described in section 2.2; (*iii*) specialized classification tasks, aimed at calculating additional features through intermediate steps; and (*iv*) final automatic classification, based on different machine learning algorithms. Features are extracted by these actors in both the initialization stage, for creating the training set, and the online operation stage, for evaluating streaming content. Three final classification algorithms have been included: Naive Bayes (NB), Sequential Minimal Optimization (SMO) and Random Forest (RF). These algorithms are described in section 4. The system can also be easily configured to encapsulate any other classification algorithm. As an additional feature, it is also possible to create an online learning loop, thus periodically feeding the training set with newly automatically classified instances, above a certain threshold of confidence [51].

In particular, for the SENT group, some features are obtained through some automatic classifiers that are implemented by few specialized actors integrated into TrollPacifier. One subsystem is dedicated to emotion detection and is built as a hierarchy of classifiers. Another subsystem is dedicated to evaluating the "abusiveness" of a text, through an ad-hoc trained classifier. The role and structure of both subsystems are described in the two following subsections.

3.5. Subsystem for emotion evaluation

A subsystem of TrollPacifier is dedicated to the evaluation of the main emotion expressed in a tweet. This classifier is effectively organized in a three-level

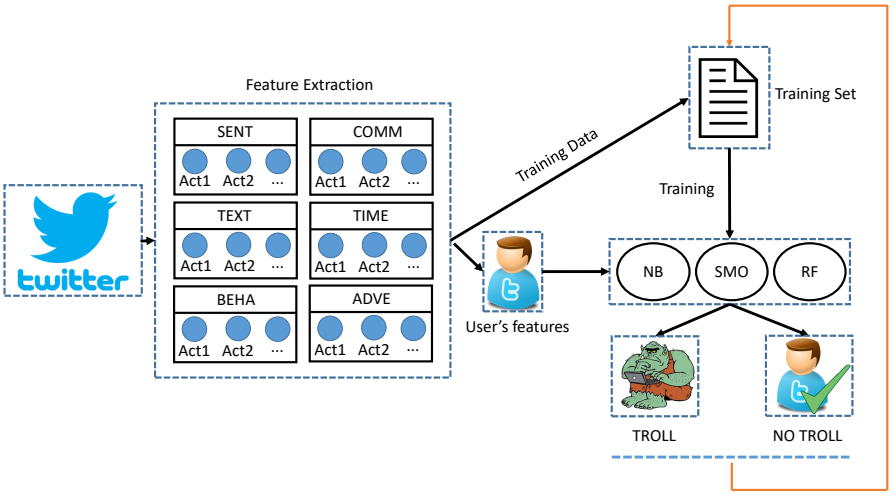

Figure 2: Representation of the actor-based system architecture.

hierarchy of four specialized classifiers, which reflect a priori relationships between the target emotions. In fact, a common approach to sentiment analysis includes two main classification stages,

1. Subdivision of texts according to the principles of objectivity/subjectivity. An objective assertion only shows some truth and facts about the world, while a subjective proposition expresses the author's attitude toward the subject of the discussion.
2. Determination of the polarity of the text. If a text is classified as subjective, it is regarded as expressing feelings of a certain polarity (positive or negative).

Extending this basic model, our subsystem adds two classifiers as an additional level for specifying the emotions which characterize subjective tweets, based on Parrott's socio-psychological model [49]. According to it, all human feelings are divided into six major states: three positive (love, joy, surprise), and three negative (fear, sadness, anger). These emotions are analyzed separately by two distinct ternary classifiers, as shown in Figure 3.

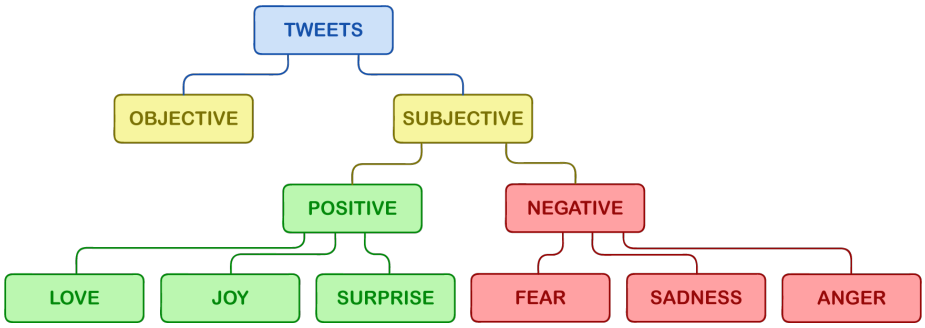

Figure 3: Hierarchical emotion classification.

Previous works show that the a priori domain knowledge embedded into this kind of hierarchical classifier makes it significantly more accurate than a corresponding 7-output flat classifier [31].

3.6. Subsystem for abusiveness evaluation

In this section we describe the ad-hoc text-based classifier for the evaluation of the overall "abusiveness" of a text. With the term abusiveness we refer to an online behavior characterized by improper or wrongful use, provoking others to finally report the author as a troll. The classifier is trained with two classes of texts: those written by alleged trolls and those written by normal users. In particular, the messages used for training the ad-hoc classifier are exactly all the post of the 6000 users described in Section 3.2. As a first step, the collected 542676 posts have been cleaned following the techniques used in [52] (text conversion to lowercase, white space stripping, Stemming, English stop words removal etc.), in order to increase the final accuracy. The training set is balanced (same number of troll posts and non-troll posts).

Since the classification of text documents requires a proper text representation, in our work we have chosen the bag-of-words model, which is simply the extraction of all words of the corpus and the representation of each sentence as the vector of the corresponding occurrences.

Due to the large number of features of the bag-of-words model (corresponding to the number of different words used in the corpus), it is necessary to reduce the model complexity, retaining only the most discriminant features. We select features according to the Information Gain (IG) criterion [50], to improve accuracy and reduce the time required by the learning step. Information Gain computes the expected entropy reduction by measuring the amount of a priori information about the class prediction when the only information available is the presence of a feature and its corresponding class distribution.

In our work, we compute the Information Gain for each feature. Then, we rank the IG scores of all the attributes in descending order; finally, in the learning step, we consider only the top k. It is to be noticed that the k value has been optimized with a grid search optimization method [53].

Grid search is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm and it is usually guided by a performance metric (in our case the classification accuracy). After the grid search optimization process, the best value found for k is about 30000 (number of features).

It is to be noticed that the previously described methods (bag-of-words model, information gain, grid search optimization, etc.) are standard approaches for the creation of an automatic text classification system [54].

The created classification model is then used for the evaluation of the "abusiveness" feature, which corresponds to the percentage of troll posts published by a user with respect to the total number of his posts.

4. Results

The experimental results described in this section show the importance of the considered features for the automatic detection of troll users. The results are presented in three separate sections, in order to highlight the effectiveness of the six considered groups of features (COMM, TEXT, BEHA, SENT, TIME, ADVE), the contribution of each feature individually, and the execution time.

Table 1: Accuracy, F-measure, Kappa statistic, AUC (Area Under the Receiver Operating Characteristic curve) and Recall for each dataset, using different Machine Learning algorithms (SMO, NB, RF).

	Accuracy (%)			**F-measure**			**Kappa**			**AUC**			**Recall**		
	SMO	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**
SENT	78.80	72.31	78.97	0.79	0.71	0.79	0.58	0.45	0.58	0.79	0.79	0.87	0.77	0.67	0.79
TIME	67.84	61.28	75.65	0.57	0.41	0.75	0.36	0.23	0.51	0.68	0.77	0.83	0.44	0.27	0.74
TEXT	67.56	64.97	68.47	0.66	0.63	0.69	0.35	0.30	0.37	0.68	0.69	0.74	0.64	0.59	0.70
BEHA	75.88	58.62	79.59	0.75	0.70	0.80	0.52	0.17	0.59	0.76	0.82	0.87	0.70	0.97	0.79
COMM	80.45	74.96	83.16	0.80	0.72	0.83	0.61	0.50	0.66	0.80	0.83	0.91	0.78	0.64	0.83
ADVE	78.07	71.70	85.01	0.78	0.67	0.85	0.56	0.43	0.70	0.78	0.79	0.92	0.76	0.59	0.85
TOT	95.52	80.25	88.28	0.95	0.78	0.88	0.91	0.60	0.77	0.96	0.90	0.96	0.95	0.69	0.89

We also analyzed the results obtained by considering a dataset containing all the features of the six considered groups (TOT dataset).

Regarding the classification methods, we decided to show the results of the 3 best classification algorithms [50] among those we tested:

1. Sequential minimal optimization (SMO). Training a support vector machine requires the solution of a very large Quadratic Programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop.
2. Naive Bayes (NB). A Naive Bayes classifier is based on the Bayes theorem. It is a baseline classification algorithm and it assumes the independence of the features.
3. Random Forest (RF). Random forests are an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

4.1. Comparison of groups of features

This section describes the contribution of each group of metrics.

Table 1 shows the different accuracies obtained with 10 runs of 10-folds cross validation on different datasets and different classification algorithms. We performed 10 runs of 10-folds cross validation in order to obtain more reliable results. The first six datasets (SENT, TIME, TEXT, BEHA, COMM, ADVE) are obtained from the one described in 3.2, by selecting only the features of the corresponding group. The TOT dataset is exactly the same described in 3.2 and shows the accuracy of the system by considering all the features.

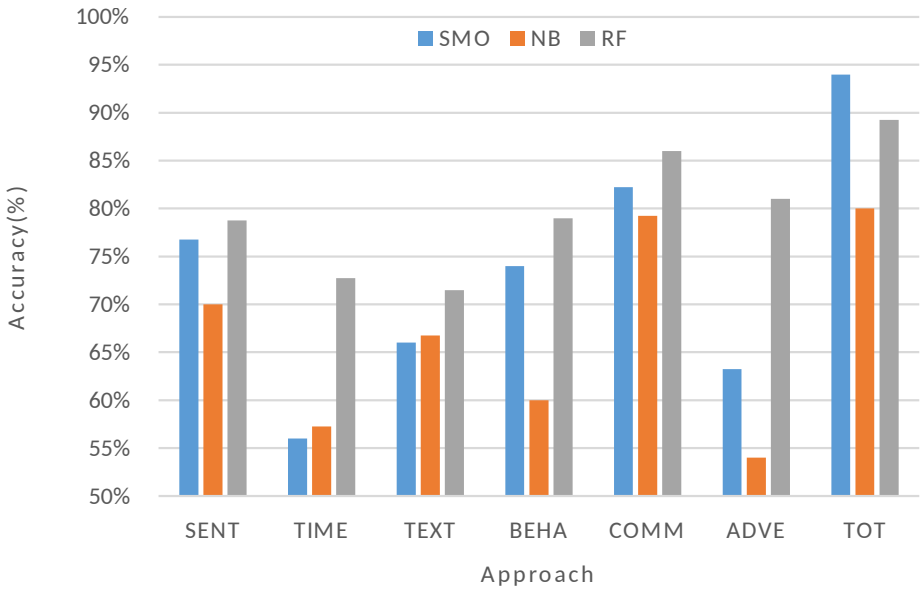

Figure 4: Accuracy obtained with different groups of features and different algorithms.

A general aspect that is deduced from the results is that some groups of metrics work better than others to distinguish the considered classes. In particular, the Community and the Advertisement group perform better than the others.

Moreover, Random Forest allows one to achieve the highest accuracy for all groups, but it is outperformed by SMO using the TOT dataset. In fact, in [55] it was demonstrated that the RF algorithm does not have high performance when dealing with high-dimensional data (like our TOT case, which clearly

Table 2: Accuracy, F-measure, Kappa statistic, AUC (Area Under the Receiver Operating Characteristic curve) and Recall obtained by removing one group of features at a time.

	Accuracy (%)			F-measure			Kappa			AUC			Recall		
	SMO	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**	**SMO**	**NB**	**RF**
TOT-SENT	94.25	79.67	87.38	0.94	0.77	0.87	0.89	0.59	0.75	0.94	0.89	0.95	0.93	0.68	0.88
TOT-TIME	95.07	82.22	89.27	0.95	0.84	0.89	0.90	0.64	0.79	0.95	0.89	0.96	0.94	0.92	0.89
TOT-TEXT	94.53	79.36	88.26	0.94	0.76	0.88	0.89	0.59	0.77	0.95	0.89	0.96	0.93	0.67	0.89
TOT-BEHA	95.05	73.87	88.98	0.95	0.67	0.89	0.90	0.48	0.78	0.95	0.90	0.96	0.94	0.54	0.90
TOT-COMM	90.40	75.51	86.92	0.90	0.71	0.87	0.81	0.51	0.74	0.90	0.88	0.95	0.89	0.59	0.87
TOT-ADVE	89.49	77.29	86.26	0.89	0.74	0.86	0.79	0.55	0.73	0.89	0.89	0.94	0.89	0.63	0.87
TOT	95.52	80.25	88.28	0.95	0.78	0.88	0.91	0.60	0.77	0.96	0.90	0.96	0.96	0.69	0.89

includes much more features than any individual group), especially in presence of dependencies.

It is also interesting to notice that Naive Bayes is often outperformed by the other two classification algorithms. Probably, this is due to the strong dependence among the features inside the same group, which are considered independent by the Naive Bayes assumption. The results can be better appreciated by looking at Figure 4.

In order to better highlight the importance of each group, we decided to evaluate complementary combinations of features. In particular, in Table 2 the first six datasets (TOT-SENT, TOT-TIME, TOT-TEXT, TOT-BEHA, TOT-COMM, TOT-ADVE) are obtained from the TOT dataset by removing the features of the corresponding group. The results are also described in Figure 5.

In addition to the evaluations shown in Table 2 and Figure 5, we have also tried to combine the contribution of each group in each classification algorithm with an ensemble learning method [56], in an effort to achieve better accuracy. The main premise of ensemble learning is that, by combining multiple models, the errors of a single classifier will likely be compensated by other classifiers, and, as a result, the overall prediction performance of the ensemble would be better than that of a single classifier. In particular, the prediction confidences of each classifier for each group have been combined using a stacking model [57] with a neural network as a meta learner (Figure 6). The hyperparameters of the neural network have been optimized using a grid search optimization method.

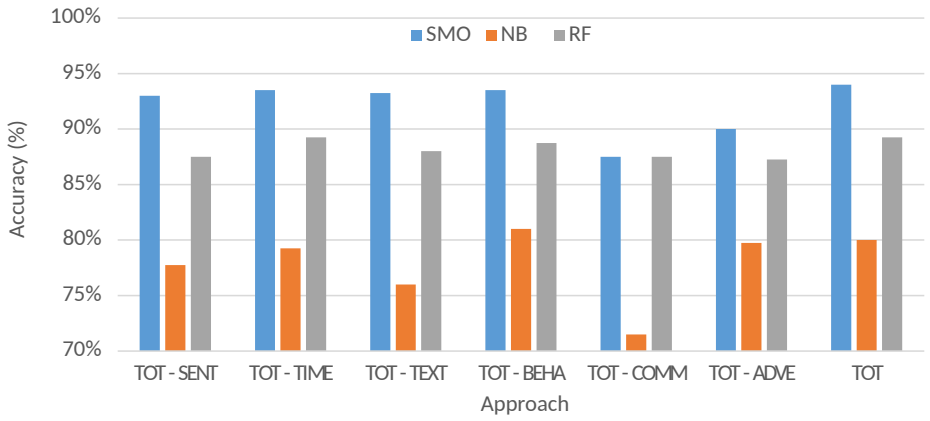

Figure 5: Results obtained by removing one group of features at a time.

In the optimal configuration, the network achieves an accuracy of 93,6%, which is lower than the accuracy obtained by the SMO classifier using all the features. This is probably due to the dependencies among features of different groups, which cannot be identified by the network since the input are only the confidence levels of previous classification algorithms.

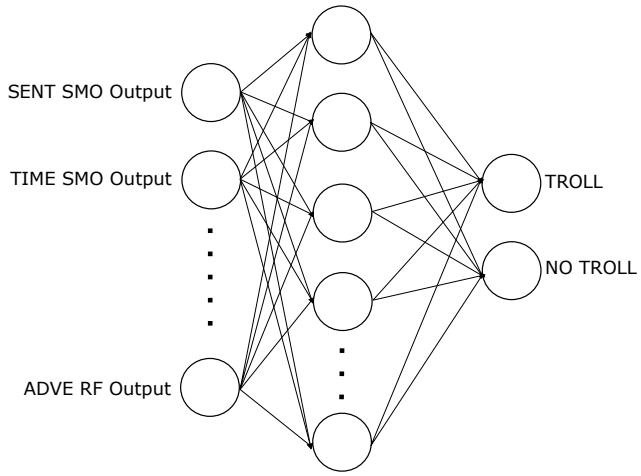

Figure 6: The neural network meta learner in the stacking ensemble learning model.

4.2. Single feature analysis and remarks

At a finer level of analysis, it is possible to assess which features have the largest influence on the results. In particular, we use the Information Gain algorithm to find the most relevant features. IG evaluates the worth of an attribute by calculating the reduction in entropy for each feature. The result is that features that perfectly discriminate the class give maximal information and unrelated features give low information.

Table 3 describes the first 10 features in decreasing order of Information Gain, together with the corresponding group.

Table 3: The first 10 features in decreasing order of Information Gain.

Description	**Group**	**IG**
#Users mentioned in the quoted tweets + #Users mentioned in the tweets	COMM	0.327
The results of an ad-hoc text-based classifier for evaluating the "abusiveness" of a text (described in 3.3)	SENT	0.275
#Answers to the user's tweets + #Retweets + #Shares	COMM	0.250
#Public lists the user belongs to	ADVE	0.235
#Followers of the user	COMM	0.228
The frequency of user's messages on Mondays from 00:00 to 04:00 (3.3)	TIME	0.139
#Urls in posts and comments	ADVE	0.191
#Replies to the user	COMM	0.178
The frequency of user's messages on Thursdays from 08:00 to 12:00 (3.3)	TIME	0.167
The frequency of user's messages on Fridays from 08:00 to 12:00 (3.3)	TIME	0.164

Features from the COMM group are the most discriminative ones (4 of the best 10 features belong to this group). However, it is to be noticed that features from different groups provide important contributions for automatic

classification and also provide useful insights about diverse aspects of online trolling. In particular, these are the most discriminative features for each group:

- COMM. The most valuable contribution, in absolute, is provided by features based on the number of mentions in the timeline. In the same group, other important features are based on the attention given to other accounts, measured on the basis of continued interactions. This indicates that troll users tend to engage in multiple and long conversations, probably due to prolonged arguments with other users. The number of followers is also discriminative, as a troll user is generally not well received by the community. The typical low level of success also leads to fewer tweets that are reshared or liked by other users.

- SENT. Among the features measuring sentiments and emotions of tweets, abusiveness is the most discriminative. In fact, it is based on an automatic classifier trained with messages written by users reported to the support channels of Twitter. This means that the lexicon used by trolls is quite distinguishing. Other features in this group provide less important contributions. The fact that trolls are not strongly characterized by emotions can be a manifestation of their Machiavellianism, which is associated with the personality of online trolls [7].

- ADVE. Generally, a troll has little incentive to subscribe to lists on Twitter, which are mainly used to remain informed on a specific topic. Instead a troll tend to publish more urls and to reshare more tweets from various sources, indicating that some trolls may be effectively engaged in various types of campaigns. They also use more hashtags, possibly to gain visibility and because they deal with multiple and diverse topics, thus lacking focus.

- TIME. It is quite interesting that the simple analysis based only on the daily and hourly frequency of messages provide quite good results. In fact, a troll produce many more tweets than a normal account, in particular

deep in the night. This can be related to availability of time and to prolonged arguments, but it can also be related to personality traits of online trolls which would deserve further studies [58].

- BEHA. While patterns of behavior are generally useful for bot detection, instead they provide minor gains for troll detection. Some features in this group, based on the number of replies to other tweets and other users, indicate an attitude of trolls to follow and engage in multiple conversations. In fact, triggering conflicts with other users result in verbal crossfires that go longer than a normal conversation.

- TEXT. Among the metrics based on the text of the tweets, the most discriminative are related to the indices of readability. Our study confirms that troll users tend to write less readable posts, as they pose less care in the drafting of their texts [22]. Other relevant features in this group include the use of emoticons, the richness of vocabulary and the number of hapaxes, i.e., words appearing only once in a user's tweets.

4.3. Execution time

Finally, to evaluate the applicability of the proposed system in real contexts, we have measured the execution time for both downloading and analyzing data. In particular, for downloading the tweets to analyze, the average time required, by user, is 1.748 s, with a standard deviation of 0.298 s. Instead, for analyzing data and then providing a user's actual features, the average time required is 43.819 s, with a standard deviation of 40.921 s. These aggregated results have been obtained from tests executed for many dozens of different users. They refer to the current implementation, which may be certainly improved through optimization and parallelization, running on a desktop PC with an i5-4210U processor, 16 GB of ram, SSD.

After having calculated all features, the time required for actual classification is practically negligible for all evaluated algorithms. In fact, the average value is 3 ms, with a standard deviation of 6 ms. To highlight some differences among the

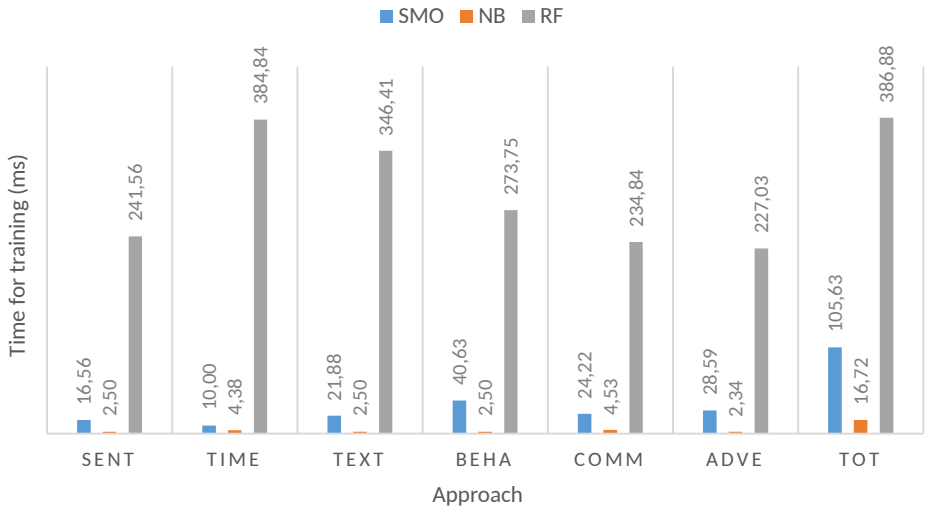

Figure 7: Time required for training the classifiers.

classification algorithms, we have also evaluated the time required for training, with results shown in Figure 7. It is worth underlying that the training process happens essentially offline, after acquiring the training set and before starting the system. However, this evaluation can be useful in the realization of more adaptable systems. In fact, in TrollPacifier, it is also possible to collect new training data at runtime to perform online learning, *i.e.*, (*i*) enrich the training set with some new instances observed while running the system, and then (*ii*) periodically update the classification model, by repeating the training process. In this case, the different computational weight of the training phase can also be taken into account.

5. Conclusion

Studying currently available research results, it can be said that the identification of troll users is possible. Some of these techniques are described as able to obtain significant results, but usually in much smaller and controllable environments than the one we have chosen. In fact, also in a large and dynamic

context like Twitter, we verified the applicability of some techniques described in the scientific literature. However, it is also evident that currently exploited methodologies can be greatly improved, since many works rely only on specific aspects of users' online presence. The fusion of different types of metrics is possible and desirable, since the problem of troll detection is complex by its nature, as it is characterized by a strong subjectivity of the act. The interest from the scientific community to the phenomenon of trolls and their automatic identification has come only recently. Rightly, the first step was to assess the applicability of proposed approaches in simpler contexts, before dealing with the larger networks. So, our own intent has been to adapt and implement known methods and new ideas for this larger context. Considering that the dimensions along which the online trolling phenomenon develops are numerous and very varied, we have been able to provide some methodological and practical guidelines. In particular, we have started applying our methodology to Twitter, as a very popular microblogging platform. The metrics and the algorithms are especially tailored for this platform. In future, we plan to extend our research work to different scenarios. We believe that this work poses good basis for a more comprehensive understanding of the problem and the value of its many faceted aspects, for building useful automatic classification tools and thus improving the conditions for more participatory online communities.

References

[1] B. A. Coles, M. West, Trolling the trolls: Online forum users constructions of the nature and properties of trolling, Computers in Human Behavior 60 (2016) 233–244.

[2] S. Rosenbaum, Is twitter toxic? can social media be tamed?, Forbes 2016 (Sep 9).

[3] C. Hardaker, Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions (2010).

[4] A. Dance, Communication: Antisocial media, Nature 543 (7644) (2017) 275–277.

[5] J. S. Donath, et al., Identity and deception in the virtual community, Communities in cyberspace 1996 (1999) 29–59.

[6] B. Kirman, C. Lineham, S. Lawson, Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls, in: CHI'12 Extended Abstracts on Human Factors in Computing Systems, ACM, 2012, pp. 121–130.

[7] E. E. Buckels, P. D. Trapnell, D. L. Paulhus, Trolls just want to have fun, Personality and individual Differences 67 (2014) 97–102.

[8] O. Bogolyubova, P. Panicheva, R. Tikhonov, V. Ivanov, Y. Ledovaya, Dark personalities on facebook: Harmful online behaviors and language, Computers in Human Behavior 78 (2018) 151–159.

[9] L. Morrissey, Trolling is a art: Towards a schematic classification of intention in internet trolling, Tech. rep., Griffith Working Papers in Pragmatics and Intercultural Communications, 3 (2) (2010).

[10] S. L. Buglass, J. F. Binder, L. R. Betts, J. D. Underwood, Looking for trouble: A multilevel analysis of disagreeable contacts in online social networks, Computers in Human Behavior 70 (2017) 234–243.

[11] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, J. Leskovec, Anyone can become a troll: Causes of trolling behavior in online discussions, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, ACM, 2017, pp. 1217–1230.

[12] B. Pfaffenberger, " if i want it, it's ok": Usenet and the (outer) limits of free speech, The Information Society 12 (4) (1996) 365–386.

[13] S. Herring, K. Job-Sluder, R. Scheckler, S. Barab, Searching for safety online: Managing" trolling" in a feminist forum, The Information Society 18 (5) (2002) 371–384.

[14] E. Cambria, P. Chandra, A. Sharma, A. Hussain, Do not feel the trolls, in: CEUR Workshop Proceedings, Vol. 664, 2010, pp. 1–12.

[15] J. Synnott, A. Coulias, M. Ioannou, Online trolling: The case of madeleine mccann, Computers in Human Behavior 71 (2017) 70–78.

[16] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, P. G. Bringas, Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying, Logic Journal of the IGPL 24 (1) (2016) 42–53.

[17] M. Golf-Papez, E. Veer, Dont feed the trolling: rethinking how online trolling is being defined and combated, Journal of Marketing Management 33 (15-16) (2017) 1336–1354.

[18] L. Derczynski, K. Bontcheva, Pheme: Veracity in digital social networks, in: UMAP Workshops, 2014, pp. 1–4.

[19] C. Dellarocas, Strategic manipulation of internet opinion forums: Implications for consumers and firms, Management science 52 (10) (2006) 1577–1593.

[20] G. King, J. Pan, M. E. Roberts, How the chinese government fabricates social media posts for strategic distraction, not engaged argument, American Political Science Review 111 (3) (2017) 484–501.

[21] F. J. Ortega, J. A. Troyano, F. L. Cruz, C. G. Vallejo, F. EnríQuez, Propagation of trust and distrust for the detection of trolls in a social network, Computer Networks 56 (12) (2012) 2884–2895.

[22] J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec, Antisocial behavior in online discussion communities, arXiv preprint arXiv:1504.00680.

[23] I. O. Dlala, D. Attiaoui, A. Martin, B. B. Yaghlane, Trolls identification within an uncertain framework, in: Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on, IEEE, 2014, pp. 1011–1015.

[24] T. Mihaylov, G. Georgiev, P. Nakov, Finding opinion manipulation trolls in news community forums., in: CoNLL, 2015, pp. 310–314.

[25] T. Mihaylov, P. Nakov, Hunting for troll comments in news community forums, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 2, 2016, pp. 399–405.

[26] A. Younus, M. A. Qureshi, M. Saeed, N. Touheed, C. O'Riordan, G. Pasi, Election trolling: analyzing sentiment in tweets during pakistan elections 2013, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 411–412.

[27] A. Lökk, J. Hallman, Viability of sentiment analysis for troll detection on twitter: A comparative study between the naive bayes and maximum entropy algorithms (2016).

[28] C. H. E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014, pp. 216–225.

[29] S. Dollberg, The metadata troll detector, Tech. Rep. Semester Thesis.

[30] C. W. Seah, H. L. Chieu, K. M. A. Chai, L.-N. Teow, L. W. Yeong, Troll detection by domain-adapting sentiment analysis, in: Information Fusion (Fusion), 2015 18th International Conference on, IEEE, 2015, pp. 792–799.

[31] G. Angiani, S. Cagnoni, N. Chuzhikova, P. Fornacciari, M. Mordonini, M. Tomaiuolo, Flat and hierarchical classifiers for detecting emotion in tweets, in: AI* IA 2016 Advances in Artificial Intelligence, Springer, 2016, pp. 51–64.

[32] N. Diakopoulos, M. Naaman, Towards quality discourse in online news comments, in: Proceedings of the ACM 2011 conference on Computer supported cooperative work, ACM, 2011, pp. 133–142.

[33] J. de-la Pena-Sordo, I. Santos, I. Pastor-López, P. G. Bringas, Filtering trolling comments through collective classification, in: International Conference on Network and System Security, Springer, 2013, pp. 707–713.

[34] H. Gao, J. Zhu, C. Li, The analysis of uncertainty of network security risk assessment using dempster-shafer theory, in: Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on, IEEE, 2008, pp. 754–759.

[35] F. Riquelme, P. González-Cantergiani, Measuring user influence on twitter: A survey, Information Processing & Management 52 (5) (2016) 949–975.

[36] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, Dna-inspired online behavioral modeling and its application to spambot detection, IEEE Intelligent Systems 31 (5) (2016) 58–64.

[37] M. Clément, M. J. Guitton, Interacting with bots online: Users reactions to actions of automated programs in wikipedia, Computers in Human Behavior 50 (2015) 66–75.

[38] J. Kunegis, A. Lommatzsch, C. Bauckhage, The slashdot zoo: mining a social network with negative edges, in: Proceedings of the 18th international conference on World wide web, ACM, 2009, pp. 741–750.

[39] S. Kumar, F. Spezzano, V. Subrahmanian, Accurately detecting trolls in slashdot zoo via decluttering, in: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, IEEE, 2014, pp. 188–195.

[40] J. Aro, The cyberspace war: propaganda and trolling as warfare tools, European View 15 (1) (2016) 121–132. doi:10.1007/s12290-016-0395-5.

[41] M. Jaitner, Exercising power in social media, The fog of cyber defence (2013) 57.

[42] J. Cordi, Social Media Revolution: Political and Security Implications, NATO Parliamentary Assembly, 2017.

[43] F. Bergenti, A. Poggi, M. Tomaiuolo, An actor based software framework for scalable applications, Lecture Notes in Computer Science (LNCS) 8729 (2015) 26–35, proc. 7th International Conference on Internet and Distributed Computing Systems (IDCS 2014); Calabria; Italy; 2014-09-22/24 [MT]. doi:10.1007/978-3-319-11692-1_3.

[44] S. Gallardo-Vera, E. Nava-Lara, Developing collaborative applications with actors, in: Proceedings of the World Congress on Engineering and Computer Science, Vol. 1, 2015, pp. 1–5.

[45] K. Chodorow, MongoDB: the definitive guide, " O'Reilly Media, Inc.", 2013.

[46] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford 1 (12) (2009) 1–6.

[47] S. Cagnoni, P. Fornacciari, J. Kavaja, M. Mordonini, A. Poggi, A. Solimeo, M. Tomaiuolo, Automatic creation of a large and polished training set for sentiment analysis on twitter, in: International Workshop on Machine Learning, Optimization, and Big Data, Springer, 2017, pp. 146–157.

[48] E. Cambria, S. Poria, D. Hazarika, K. Kwok, Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings, in: AAAI, 2018, pp. 1795–1802.

[49] W. G. Parrott, Emotions in social psychology: Essential readings, Psychology Press, 2001.

[50] R. S. Michalski, J. G. Carbonell, T. M. Mitchell, Machine learning: An artificial intelligence approach, Springer Science & Business Media, 2013.

[51] P. Fornacciari, M. Mordonini, A. Poggi, M. Tomaiuolo, Software actors for continuous social media analysis, CEUR Workshop Proceedings 1867 (2017) 84–89.

[52] G. Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, S. Manicardi, A comparison between preprocessing techniques for sentiment analysis in twitter., in: KDWeb, 2016, pp. 1–11.

[53] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[54] Y. Yang, J. O. Pedersen, A comparative study on feature selection in text categorization, in: Icml, Vol. 97, 1997, pp. 412–420.

[55] T.-N. Do, P. Lenca, S. Lallich, N.-K. Pham, Classifying very-high-dimensional data with random forests of oblique decision trees, in: Advances in knowledge discovery and management, Springer, 2010, pp. 39–55.

[56] M. Sewell, Ensemble learning, UCL Research Note 11 (02) (2011) 1–12.

[57] G. Wang, J. Hao, J. Ma, H. Jiang, A comparative assessment of ensemble learning for credit scoring, Expert systems with applications 38 (1) (2011) 223–230.

[58] P. K. Jonason, A. Jones, M. Lyons, Creatures of the night: Chronotypes and the dark triad traits, Personality and Individual Differences 55 (5) (2013) 538 – 541. doi:https://doi.org/10.1016/j.paid.2013.05.001.