future internet

MDPI

Review

A Survey on Troll Detection

Michele Tomaiuolo, Gianfranco Lombardo, Monica Mordonini, Stefano Cagnoni and Agostino Poggi *

Department of Engineering and Architecture, University of Parma, 43124 Parma, Italy; michele.tomaiuolo@unipr.it (M.T.); gianfranco.lombardo@unipr.it (G.L.); monica.mordonini@unipr.it (M.M.); stefano.cagnoni@unipr.it (S.C.)

* Correspondence: agostino.poggi@unipr.it; Tel.: +39–0521–905728

Received: 17 January 2020; Accepted: 7 February 2020; Published: 10 February 2020

Abstract: A troll is usually defined as somebody who provokes and offends people to make them angry, who wants to dominate any discussion or who tries to manipulate people's opinions. The problems caused by such persons have increased with the diffusion of social media. Therefore, on the one hand, press bodies and magazines have begun to address the issue and to write articles about the phenomenon and its related problems while, on the other hand, universities and research centres have begun to study the features characterizing trolls and to look for solutions for their identification. This survey aims at introducing the main researches dedicated to the description of trolls and to the study and experimentation of methods for their detection.

Keywords: troll detection; antisocial behaviour; social media

1. Introduction

The extension and pervasiveness that the Internet has reached in recent years has led to the emergence of many platforms specialized in communication services. Social media have been adopted in many different countries by the public [1] as well as by companies [2–4] In addition, "being social", in contrast to "being a troll", has been shown to be very important for the quality of human interaction in the digital sphere; this attitude can be assessed in different ways [5–7].

A troll is an individual with an antisocial behaviour that incites other users acting within the same social network [8]. In particular, a troll often uses an aggressive or offensive language and has the purpose to slow down the normal evolution of an online discussion and possibly to interrupt it [9]. Only recently has it been possible to pay proper attention to this problem, so that many renowned press bodies and magazines have started 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, favoured by the increasing occurrence of behaviours like the one described above.

This kind of behaviour is not fully characterised and, up to now, it has been difficult to find an accurate description for the word "troll", since the act of trolling is strongly subjective. The lack of an agreed-on definition for the term "troll" has resulted in poor comprehension and in low interest for the research community. The need for dealing with this problem has therefore emerged over time, along with studies conducted by several universities and research centres.

After removing applications which are not strictly related to the main topics taken into consideration (social sciences, computer science and engineering), Scopus, as of February 4, 2020, lists 636 papers having the term "troll" in the title or abstract or as a keyword, when limiting the search to those three subject areas, 401 of which are related with the two latter topics, and 192 only to "computer science". Adding the keyword "detection" brings the total down to 51 papers, whose distribution in time shows a clear increment after 2015. Even when limiting the search to this rather narrow "topic", it is quite clear that the recent interest in this kind of application has been stimulated

by recent events having worldwide resonance. For instance, the hype arisen by the alleged influence of Russia in the latest United States presidential elections has shown that malicious behaviours in social networks are not only a "local" menace within limited groups of users of a social network, but can assume world-wide dimensions, up to affecting world politics [10–13].

This article presents a short introduction to the problem of trolls and to their detection. Its goal is by no means to provide an exhaustive overview of the approaches to troll detection but to list and discuss a number of paradigmatic examples which can be a good starting point for those who are interested in joining the increasing community of researchers on this topic and providing their own contribution. As such, it has not been originated by a systematic and quantitative "search and refine" process. It is rather the result of a top-down incremental browsing of the literature, starting from seminal papers and iteratively following new paths according to the suggestions and discoveries of the newly accessed ones. Substantially, it reflects and expands the strictly application-oriented analysis of the state-of-the-art we performed in the first stage of development of [14].

The paper is organised as follows: the next section presents the trolling phenomenon. Section 2 discusses the main methods used to detect trolls. Finally, Section 4 concludes the article by discussing the most relevant open problems with troll detection and by suggesting directions for future work.

2. Social Media and the Trolling Phenomenon

With the rise of social media, users of these services have been able to benefit from a new simple and fast way of communicating, capable of connecting separate individuals both physically and temporally. The Computer-Mediated Communication (CMC), however, cannot fully replicate the dynamics of verbal communication since it is exclusively based on sentences constructed very simply. This increases the chances of misunderstandings and, consequently, of conflicts. CMC can also provide various degrees of anonymity, which can induce users to feel a sense of freedom and immunity from being held accountable for inappropriate behaviours. As a result, there has been a change in the way of communicating that has enabled users to limit the amount of personal information revealed. This has led to the development of widespread trolling within the CMCs. A troll has been defined as an individual who displays a negative online behaviour [8], or as a user who initially pretends to be a legitimate participant, but later attempts to disrupt the community, not in a blatant way, but aiming to attract the maximum number of responses from the other community members [15] Trolls are also described as individuals who derive pleasure from annoying others [16] and, in fact, recent works have discovered that sadism is closely associated with those who have trolling tendencies [17]. The etymology of the word may come from a particular practice used in fishing, where a fishing line is dragged behind the boat in order to catch fish, while other sources trace the origin of the term to the monsters of Norse mythology, who used to wait under the bridges to ambush unsuspecting passers. The first references to the use of the word "troll" on the Internet can be traced back to Usenet, a forum community popular in the eighties.

2.1. Troll Definition and Features

A troll's goal is to make fun of a person; if this is analysed as a pragmatic act, it may be divided into three basic components: (*i*) a pseudo-intention, (*ii*) a real intention and (*iii*) a stimulus. The naivest users can identify only the fictitious intention, the more experienced ones can also identify the actual purpose, correctly recognizing the action of the troll [18].

In sociology, the term has become a synonymous for all negative behaviours that can be found online, but it is necessary to define each one more precisely, in order to understand and discuss these behaviours in an academic way. Hardaker [8], studying the behaviour of some users within a social media context, has found that the act of trolling is manifested in four interrelated ways:

- *Deception*: within a community like Usenet, as in any other question-and-answer (Q&A) forum, if a troll wants to have some chances of success, he must keep his real intent of trolling hidden. He will attempt to disrupt the group, trying to stay undercover. In fact, it is not possible to determine with certainty whether someone is causing problems intentionally and to label that person as a troll, because it may be simply a novice user or a discordant voice. Perhaps it is

easier (for a user or for a supervisor) to identify ambiguous behaviours and then to assess whether they are maintained over time. An example may be the "pseudo-naïve" behaviour, that occurs when a troll intentionally disseminates false or inaccurate advice or pretends to ask for help, to provoke an emotional response in the other group members [15].

- *Aggression*: a troll who is aiming at generating a conflict can use a provocative tone towards other users. These are malicious or aggressive behaviours undertaken with the sole purpose to annoy or provoke others, using ridiculous rants, personal insults, offensive language or attempts to hijack the conversation onto a different topic.
- *Disruption*: it is the act of causing a degradation of the conversation without necessarily attacking a specific individual. A behaviour of this type includes sending senseless, irrelevant or repetitive messages aimed at seeking attention. This has also been referred to as trolling spam, linked to the common spam, but separate from it, as it is driven by the intention to provoke negative responses.
- *Success*: one of the most curious aspects of the problem is that, often, a troll is acclaimed by users for his success both in relation to the quality of his own joke, i.e., for being funny, and for the way others react to it. In fact, some responses to the provocation—whether they are angry, shocked or curious—are regarded as a "bait" to the troll's joke or, in other words, a demonstration that those who are responding were unwittingly duped by the pseudo-intent of the troll without being aware of the troll's real goal. The attention of the group is similarly drawn even when the quality of a troll's joke is low and everybody can understand his real intent or when an experienced user can respond to a troll's message in a manner that prevents him from falling into the prepared trap, possibly trying to unnerve the troll. So trolling, despite being a nuisance for users, may end up being the centre of attention of the group for its real purpose and not for his pseudo-intent. Therefore, this aspect is related to how the group reacts to the troll and not to its modalities.

It is clear that trolling is a more complex problem than just provocative attacks. Although the concept may seem to be tied to the meaning of some words like rudeness, arrogance, impertinence and vulgarity, these do not provide an accurate description of the troll's attitude since, typically, trolling consists in keeping hidden the real intent of causing problems. In addition, in communities in which users are less vulnerable, more experienced or emotionally detached, the phenomenon can be seen as a playful action. As a result of this analysis, Hardaker provides an academic definition:

"A troll is a CMC user who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement. Just like malicious impoliteness, trolling can (*i*) be frustrated if users correctly interpret an intent to troll, but are not provoked into responding, (*ii*) be thwarted, if users correctly interpret an intent to troll, but counter in such a way as to curtail or neutralize the success of the troll, *(iii)* fail, if users do not correctly interpret an intent to troll and are not provoked by the troll, or, (*iv*) succeed, if users are deceived into believing the troll's pseudo-intention(s), and are provoked into responding sincerely. Finally, users can mock troll. That is, they may undertake what appears to be trolling with the aim of enhancing or increasing affect, or group cohesion" [8].

2.2. Troll's Damages

Inexperienced or vulnerable users of the Internet communities who trust trolls, are involved emotionally, or communicate private information, may feel trolling particularly painful, distressing and inexplicable; given the distributed and asynchronous nature of online discussions, this may have consequences in the long term. These practices, although clearly problematic, are common and often tolerated, in part because libertine values widespread on the Internet consider insulting speech as a manifestation of freedom of expression [19]. In extreme cases, malicious users use the CMCs to commit crimes such as defamation, intake of others' identities or cyberbullying. To counteract this, some online communities implement identity verification processes and disable the options that allow simultaneous communication between users [20]. Nevertheless, recently, the propensity to

trolling seems to be widespread, which is alarming many of the most important social networks because, in extreme cases, it has led some adolescents, like Amanda Todd, to commit suicide [21]. These attacks are usually directed not only to individuals, but to whole communities. For example, a growing number of tribute pages on Facebook are being targeted, including one in memory of the victims of the shootings in Cumbria and one dedicated to soldiers who died in the war in Afghanistan [22].

Even when trolling does not come as a direct attack, it can still be a threat because it can manifest itself in subtler ways, for example as a mean to try to manipulate others' opinions. In fact, the rise of the Internet has allowed companies, organizations and governments to freely disseminate false rumours, misinforming and speculating, and to use other dishonest practices to polarize opinions [23] It has been shown that the opinion of a user on certain products or on politics can be influenced by the comments of other users [24] This way, gaining popularity is made easier for companies and for those political parties which make use of reputation management services, i.e., people paid to hijack the opinions on their behalf. There are many publications that describe how these behaviours have a strong impact on current events. For example, [25] says that in China there is an "army" of two million people, daily active on social networks, who flood citizens' digital debates with comments and opinions that lead those discussions towards more acceptable topics, preferred by Beijing government.

The danger of social media abuses in the political sphere, with the aim of eventually affecting important election and voting events, is raising great concerns. One of the events that have recently attracted wide attention is the foreign interference during the 2016 US Presidential election. In particular, Russia has been accused by the US Congress of conducting a systematic mass manipulation of the public opinion, using both human operators and software controlled accounts (i.e., bots). The accusation has been accompanied with a list of 2752 Twitter accounts, allegedly tied with the "Internet Research Agency" (IRA), described as a "troll farm" based in Russia [26]. This is one of the first revealed large scale organizations, systematically using human operators for political propaganda and deceptive interference campaigns. The accounts in the list used to spread politically-biased information and have been later deactivated by Twitter. The list provided by the US Congress is being used in a number of research works aiming at the automatic detection of online trolls, particularly those interfering in the political sphere [10–12]. Some studies deal, in particular, with the diffusion of false information and fake news by trolls, describing the phenomenon as "disinformation warfare" [13]. In this sense, the problem can also be analysed from a general viewpoint of information quality assessment applied to social media [27].

However, even without considering these extreme consequences, trolling remains a vexing problem because, even when undertaken in an innocent way, 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. With such a high number of degrees of freedom of expression, the exclusion of individuals with an unpleasant behaviour as trolls needs to be considered very carefully, since it can trigger side effects. Nevertheless, it is appropriate to avoid many undesired disruptions and maintain clean and focused discussion threads.

2.3. Coping with Trolls

As a result of the complexity of modern social media, identifying and banning trolls is a daunting task. It is important to make users aware of trolling behaviours, but these kinds of warnings do not reduce the phenomenon. A common way to neutralize a troll is to avoid speaking to him; to this purpose, expert users advise novices not to "feed" him ("Do not feed the troll!"). Usually, online communities comprise a special group of users known as moderators, who have the ability and responsibility to ban those users who are considered malicious. The main benefit of this entirely supervised approach is the possibility to determine if a user is a troll, based on his actions or his comments on the network. This solution has two major drawbacks: subjectivity and scalability. A supervised system can be reliable in medium-sized networks or in small ones, but does not scale to social networks with a large number of members and contents. Some of these communities have tried

to overcome these limitations by delegating some control functions to its users. To do so, they have been provided with a mechanism to point out offensive contents or behaviours to moderators. Subsequently, it is up to moderators to determine if the reported user is a troll or not, or if a reported content has been produced by a troll or it was simply a controversial response. In addition, users may be given the opportunity to express their own opinion on others or on their comments. Other examples come from sites like Twitter or Facebook, which allow users to report a tweet or a post as inappropriate and then they automatically remove it, once it reaches a certain number of reports.

These decentralized approaches reduce the problem of subjectivity, because the decisions are delegated to all network users, and the one of scalability, since there is no authority that operates as a controller [28]. However, even by assigning the control task to different persons, it is possible to come upon other limitations generated by the "manual" countermeasure taken in response to the problem. This happens because, generally, the community cannot manage to react instantaneously to the report and the result may be too slow; consequently, the users are not protected from being emotionally hurt by a troll. In view of all these things, it is possible to claim that the best solution is to automatize the process by implementing a troll detection algorithm. Historically, when trolling was still unknown and had not been studied in detail yet, the researchers' focus was on a phenomenon already known and similar to the case taken into exam: the so-called "fake", that occurs when a user falsifies his identity.

Occasionally, this kind of user turns out to be a troll [21]. In fact, the ability to create an online *alter ego* allows the trolls to publish news, reviews or multimedia material intended to discredit or attack other people that may or may not be aware of the attack. Starting from these observations, still mostly valid, new anti-trolling methods have been implemented over the years, mainly consisting in the identification of accounts that use the same IP address and in blocking the fake ones. These methods are based on spotting names and unusual activities like, for example, the dispatch of many messages to non-friend users or the high rate of declined friendship requests [22] More recently, the solutions considered to solve the troll detection problem rely on artificial intelligence and data mining approaches, ranging from sentiment analysis to social network analysis [28].

3. Troll Detection Methods

Several research directions have been considered by researchers in order to solve the troll problem.

The first to be taken into consideration was to automatically analyse online contents through a natural language processing (NLP) approach. A rudimentary NLP technique involves the calculation of the negative words contained in a given post, in order to measure the comment's degree of hostility. Another approach is based on the subdivision of the content of the post in n-grams, i.e., sequences of n elements included in the text (in this case, words or characters, but also emoticons). These elements are compared to other well-known n-grams or they are used to create statistical models aimed at identifying trolls [29]. More sophisticated methods have been developed making progress in this direction; they try to spot trolls using sentiment analysis techniques, i.e., by understanding and measuring the sentiment of the text [22–32]. In fact, by attributing a certain emotion to the words in a sentence, it is possible to evaluate the predominant sentiment.

Emotions such as anger or rage are clues for detecting a troll's comment [22]. According to the authors of [29,33,34], the information acquired from single comments is not enough to perform a correct analysis and, consequently, they try to integrate methods to verify the consistency of the text according to other comments and their topic. A second research direction involves the Social Network Analysis (SNA) of the communities in order to identify possible trolls [9,28,35] Other analyses on data from users [36,37] are carried out, in order to identify users with antisocial behaviours within a community. In general, the SNA approach makes it possible to extract the information needed to assess the attitude of a user.

Finally, another approach is to combine all the previous ones by identifying heterogeneous groups of features to feed more complex machine learning models. In practice, features of trolls and legitimate users are collected, through the analysis of: writing style, sentiment, behaviours, social

interactions, linked media and publication time. A machine learning approach is finally used to identify trolls with very high accuracy [12,38,39].

A useful classification of troll detection methods is based on the type of information they use for detection. In particular, we identify four main types of information: posts, discussion threads, user behaviours and community relationships. A detailed analysis of the four methods is reported below.

3.1. Post-Based Methods

A user of a Q&A forum, an online community or a social network, expresses his opinions and his considerations through his actions within the platform, usually in a written form. A malicious individual, as well, expresses his intentions using the same means of communication. Consequently, it is possible to assume that a good method to identify a troll could consist in the analysis of the textual content and the information resulting from his comments [40] In fact, concerning this second class of methods, some of them succeed in dividing the users into legitimate and malicious ones by: (*i*) evaluating the relevance of their comments on a specially engineered scale of values, (*ii*) making use of classifiers, made with custom-made training sets, possibly with the help of crowd-sourcing services or (*iii*) evaluating their Automated Readability Index (ARI) [41] since it has been shown that a troll is more likely to write in a less comprehensible language compared to a normal user [36].

On the contrary, most approaches carry out the troll detection task by analysing the textual content, using the tools provided by Sentiment Analysis. For example, the method described in [31] is applied within the Twitter social network to identify political activists hostile to other parties and to evaluate the degree of conflict between two different factions during the 201[29] electoral period in Pakistan. The researchers use a tool called SentiStrength, very useful to estimate the "force" of a sentiment (positive or negative). It attributes a value to single words, ranging between +1 (non-negative) and +5 (extremely positive), or from -1 (non-positive) to -5 (extremely negative). Combining such values, it is then possible to evaluate the general sentiment of a sentence. Another study[31], likewise aimed at analysing political discussions on Twitter, tries to spot malicious users by analysing the content of their tweets. In particular, it tries to establish if sentiment analysis is an appropriate technique and which learning algorithm fits best between Naïve Bayes and Maximum Entropy. By examining the group of users with the lowest score from both classifiers, it is possible to find out that a lot of them are false positives, i.e., thinkers that have an unpopular but defensible point of view, while only a few users can be considered actual trolls. The results emphasize that neither method is recommended to carry out troll detection on Twitter, because, despite its rich API, it is not appropriate to obtain useful information for correctly training of the classifiers. Furthermore, the study underlines that the approaches based exclusively on sentiment analysis do not yield satisfactory performances when the aim is to determine whether a comment has been written by a malicious user.

Starting from these considerations, different paths have been explored, one of which also aims at examining data which are disconnected from the post, and will be taken into consideration in the following paragraphs. Others have scrutinized or extended methods based on sentiment analysis. Considering those limitations, a possible way forward is to enrich the analysis with measurements acquired from the comments themselves. In fact, the approaches that we have seen so far were only focused on the sentiment of the sentence but they ignored other features that can be gathered from the text. One of them, for example, is that a troll is more likely to write short comments, maybe because he writes faster replies compared to a non-malicious user that writes more elaborated and longer sentences [29]. This kind of information is important for the methods tackling the issue through the use of metadata, which may include some specific properties of the comment like, for instance, the author, the key words, the creation date, the length of the comment, the number of injurious words, etc. The work illustrated in [29] tries to detect the presence of trolls inside the reddit.com portal using solely the metadata, highlighting some characteristics according the criteria set out above. All the information gathered has been collected in attributes of instance variables used to train a Support Vector Machine (SVM) classifier, that, once tested, has yielded a good identification percentage, approximately 70%.

The results show that the approach based exclusively on metadata is less accurate than the ones based on sentiment analysis. However, a combination of the two could bring benefit to both methods as, for example, happens in [32] a work which investigates the limits of troll detection methods that only use features derived from sentiment analysis. In fact, the peculiarity of this study is that the metadata that are used are statistics directly obtained from sentiment analysis. Specifically, the following characteristics are used: the most negative comment, the most positive one and the average of all the sentiments of a post, as well as values aimed at representing the probability that a comment could contain ordinary, negative or positive terms. This information is then grouped by thread, and eventually by user, with the purpose of providing a description of the instances. Those instances will finally be passed to SVM classifiers, that will establish to which extent a certain user can be identified as a troll.

Given the limitations of these approaches, some researchers [22] have decided not to use metadata but to employ a new paradigm for text analysis focused more on semantics rather than on syntax and more inclined to understanding the sense and not the sentiment of the text, namely the "sentic computing". Additionally, it turns out to be particularly useful for understanding what is manifested implicitly, i.e., when the context and the concepts that depend on the domain become important. In fact, this model is not shaped on static learning models, but uses tools based on common sense and ontologies on a specific domain.

In particular, two tools are used to extract the semantics and the sense from online posts. The first is called "AffectiveSpace", a language display system that transforms natural language from a linguistic form into a multidimensional representation as a vector space (of size 14,301 x 117,365) in which every concept is represented by a vector within it. The second is called "The Hourglass of Emotions", and is designed to recognize, comprehend and express emotions based on Minsky's definition of emotions, according to which our mind is formed by four independent emotional spheres (pleasantness, attention, sensitivity and attitude), that turn on and off at different intensities and that are able to categorize the different levels that compose all the emotional states.

The filtering process involves four main components: (*i*) an NLP module, that carries out a first discernment of the text, (*ii*) a semantic parser, that tries to extract the key concepts and for each of them provides the related frequency in the text, (*iii*) the connotation (positive or negative) and (*iv*) the degree of intensity by which the concept is expressed. They form an "Affective Space", that takes the concepts found in the vector space, groups them in the "Hourglass" model, where the effective value of the emotional spheres is deduced according to the position that the vectors occupy in the space. A troll detector takes the "sentic vectors" from the previous step as input. It uses the information given directly by the semantic parser to eventually calculate the "trollness" and, if necessary, to block the presumed troll. In fact, the main purpose of this identification process is to take advantage of the cognitive and affective information associated to the natural language in order to define a level of "trollness" for each post. According to this evaluation, it classifies the users and prevents malicious users from hurting other individuals in their same social network.

3.2. Thread-Based Methods

Other research works attempt to bring the troll identification to a higher level of analysis, by studying not only the single comments, but entire discussion topics. This kind of hybrid approach incorporates some of the techniques described in the previous subsection, but also adds new information obtained from the context into which the messages are incorporated. Among them, [29] adopts a combination of statistical and syntactic metrics, along with other elements related to the users' opinion. Some of these measurements are similar to the ones which have been previously discussed. Others manage to summarize more general properties of the discussion, like the number of references to other comments, how many times a certain 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 computed based on the cosine similarity [36,37].

The approach conceived in [33] evaluates the problem from the same point of view, but using different concepts. It is based on the Dempster-Shafer theory, i.e., a generalization of Bayes'

probability 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 characterise the messages according to: (*i*) the senselessness of their content, (*ii*) their degree of controversy and (*iii*) whether they are on-topic or not. Thus, each message is associated with values that identify precisely these three characteristics. Thanks to these measurements and to the use of the mathematical tools made available by the above-mentioned theory, the degree of hostility is calculated throughout the messages of the different users that contribute to the thread. The data thus extracted are grouped to quantify the total conflict between a certain individual and another. For this measurement, an inclusion metric is used as a tool to quantify the conflict between two "belief functions". Finally, based on the results thus achieved, it is possible to tag the users as trolls or non-trolls. This approach has the merit of generalizing the study about the trolling phenomenon not only on single comments, but also on its context.

Moreover, the framework developed by [33] has not been designed for a particular environment and, since it is generic, it can be applied to any kind of forum. Nevertheless, the article acknowledges that different analysis methods exist to evaluate the nature of the content of the messages and that they are capable to extrapolate the useful data for determining the above-mentioned characteristics. However, this important aspect is not fully developed in the article. For this reason, the final tests on the method are carried out on synthetic data.

3.3. User-Based Methods

Other approaches observe the problem from an even more general point of view, without being limited to the information contained in users' posts, and in their threads. They are able to deduce whether a user is a troll or not by considering the overall attitude of the user within the community. In fact, often, an individual enrolled in a particular service, expresses an anti-social behaviour in most of his/her activities within the network. In this case, the malicious actions are not sporadic and limited to particular topics, but they are the result of an *a priori* attitude that characterises most of the user's online experience. Therefore, the conduct adopted by these users can be studied and catalogued in order to prevent similar attitudes, identifying them early and minimizing their effects. To this purpose, it is useful to study, analyse and characterise the conduct that trolls can take within the online community. The attention is not focused onto an individual post or onto the whole discussion thread, but onto the actions of malicious individuals and the trace of their behaviours throughout their life within the community. Consequently, the methods described in that study do not predict whether a particular comment is malicious or not, but they try to determine if a user takes a negative conduct, based on all his activities.

These data mining methodologies, in addition to a general analysis on the behaviour of users, may use certain techniques based on the study of trolls at the comment (or thread) level, through sentiment analysis or cosine similarity method. In this sense, they are considered as hybrid approaches, according to the taxonomy of models presented in this research work.

The need for integration of other metrics for the study of this problem has not been assessed empirically, but has emerged from some works available in the literature [31]. In fact, as demonstrated in the works previously described, the results of sentiment analysis must be merged with other more sensitive and specific metrics to ensure that the troll detection is effective. Otherwise, the analysis needs to be improved to take additional aspects into account.

Based on these considerations and observing the online history of an individual, some researchers have focused their efforts onto the extraction of general information about the users. The aim has been to study the most significant parameters for the characterization of a troll and obtain a better perspective of their behaviours. In particular, [37] presents a study on anti-social behaviour in online discussion communities, focusing on those users that will eventually be banned. This leads to (*i*) a characterization of malicious users, (*ii*) an investigation on the evolution of their behaviour and (*iii*) the reactions taken by the community. In order to discern respectful users from malicious ones, researchers from Stanford and Cornell rely on the community and its moderators to see who they consider an individual with a malicious behaviour. Hence, they make some aggregate analyses about

the users who will be permanently banned from the network, since these appear to be clear examples of users with an antisocial behaviour. This is therefore an approach on a large scale, and data-driven, that studies antisocial behaviour in three large communities (CNN.com, Breitbart.com and IGN.com). Over 18 months, more than 40 million posts have been collected, with the objective of obtaining quantitative specifications and to develop, as a result, tools for the early detection of trolls. The analysis focuses on the users subsequently banned by the moderators, defined as "Future-Banned Users" (FBUs), discriminating them against more respectful users, defined as "Never-Banned Users" (NBUs). By comparing the data of these two groups, one may notice that the language used in the posts by the two types of users is significantly different. Specifically, the FBUs tend to: (*i*) write comments which are more difficult to understand, (*ii*) not being inclined to keep the conversation "on-topic" and (*iii*) to use an adversarial language more frequently, i.e., they use fewer positive words in favour of other more irreverent words. In addition, the messages of FBUs focus on a few discussion threads and more rarely they comment on several topics. However, they tend to interact more frequently than others, i.e., they contribute with more posts per thread, probably resulting from the continuation of the conversation with other users. In fact, they receive more answers than average, suggesting that success in attracting attention can be synonymous with abnormal behaviours.

Furthermore, a high rate of post cancellations is an accurate indicator of unwanted behaviours, because only moderators can delete them. In fact, they act in accordance with community policies, which generally tend to consider disrespect, discrimination, insults, profanities and spam as unsuitable behaviours. Another feature, the signalling of a post, is correlated with cancellation, as the reported posts are read by moderators and possibly deleted. Another aspect that emerges is that the cancellation rate for the posts of an FBU increases over time, while that of an NBU remains relatively constant. So, we can say not only that FBUs have a worse writing style than NBUs, but also that the quality of their interventions worsens over their virtual life. This is due to the way the community considers an individual, which can play a role in his evolution into trolls. In fact, those users, who are censored or have comments removed, are more likely to behave in a worse way in the future.

Subsequently, thanks to the performed analysis, researchers have been able to develop a method suitable for identifying an FBU, which considers the first ten posts from a user [38]. It can determine the class, based on four classes of features, in which more metrics are defined:

- *Post content*: word count, readability metrics, actual content, etc.
- *User activity*: daily number of posts, maximum number of published posts in a thread, total number of posts in response to specific comments, etc.
- *Reactions of the community*: votes per post, total number of indicated posts, total number of responses, etc.
- *Moderator's actions*: number of removed comments, etc.

Some results have been obtained over multiple experiments, using different types of classifiers and different information subsets. They show that the system is able to predict, with over 80% accuracy, when an individual will be banned. The performance of the classifiers remains high, even in different domains. This is made possible by training classifiers on data from a community, but testing them on others. The moderators' actions are the most effective features to classify malicious people; among these features, the most relevant is the rate of deleted posts. The group of features derived from the behaviour of the community is the second most effective group; among them, ranked by effectiveness, are the percentage of negative votes received and the number of reported comments, while the degree of similarity of the text to the previous comments (assessed by the cosine similarity) does not improve the performance of the classifiers. The third most effective group of features includes those produced by the users' activities; among them, the single most effective feature is the number of daily posts. The post content appears to be the weakest feature, suggesting that the classifiers based only on sentiment analysis do not provide significant benefits to performance.

A similar study was carried out in [38], where a much more pragmatic approach is taken for creating the dataset: a user has a much better chance of being a troll if he is repeatedly identified as such by different users. Specifically, the authors, studying the community of an online newspaper

(Dnevnik.bg), consider trolls to be those users who are called so by a number of different individuals, with the threshold initially set at 5. Instead, non-troll users are never identified as such, with the restriction that a user must have posted at least 100 comments to be considered relevant to the analysis. The dataset contains a total of 317 trolls and 964 benign users. From all those accounts, the authors were able to extract some basic features such as: (*i*) the total number of interventions, (*ii*) the number of days spent after inclusion, (*iii*) the number of days in which the user has posted at least one comment and (*iv*) the number of threads in which he participated. Subsequently, they have derived more specific metrics (based on some assumptions documented in their article), including:

- *Rating*: percentage of user comments in each level of evaluation (very positive, positive, average, negative and very negative).
- *Consistency of the comments with respect to the topic*: cosine similarity between the comments of the same thread.
- *Order of comments*: number of times in which a user is among the first ones to comment on a discussion thread.
- *Comments most loved/hated*: number of times that a comment is among the most loved or hated in a thread (with various thresholds).
- *Answers*: number of responses to other comments, number of responses to other answers, etc. Other features are then generated by fusing these values with those based on votes.
- *Time*: number of comments made at different times of the day and on daily and weekly basis.

In total, 338 attributes are considered. From each attribute, it is possible to derive a specular measurement, in which each value is normalized by one of the above-mentioned basic features, to avoid penalizing, or possibly avoid favouring, the "younger" users in the community, obtaining, in fact, twice as many attributes.

Thanks to these data, which are normalized in [−1,1], it was possible to train a SVM classifier, testing various configurations of the parameters (for example by varying some of the thresholds) and specific sets of data, able to recognize the troll with an accuracy of 82–95%. In conclusion, both the normalized attributes and the absolute ones are useful for the analysis. In particular, the features related to the number of replies and the votes obtained by other users are very significant.

Instead, the following features are less significant: total number of user comments, order of comments at the beginning of a topic, consistency of a comment with respect to the topic. In addition, the features based on time have not been taken into account; in fact, their removal improves classification accuracy.

Although the two aforementioned works adopt different approaches to creating their training set, it is possible to note that many of their respective conclusions are conceptually similar. This proves that the distinctive features of antisocial behaviour, regardless of the community in which they are studied and the construction of the dataset, are not linked to specific environments. In this way, it is possible to infer not only the applicability of the identified features, but also the generalization to more platforms.

3.4. Community Based Methods

This is the highest level on the scale of the methods that try to find a solution to the problem of troll detection, through the study of the relationships within the online community. This is possible thanks to the use and adaptation of tools provided by social network analysis.

In the literature, the first study which explores this field is reported in [9]. The main objective of this research is to extract all the information obtainable from a social network. In fact, the identification of trolls is just a part of the study. The analysis is performed on Slashdot Zoo, a portal where any user has the option to label any other user as friend or foe. Thanks to this peculiarity, the extracted social graph has some links with negative weights (to represent correctly also the distrust connections). The basic data have then been enriched with much more information, by adapting existing social network analysis techniques to this network topology.

Some of these data are useful to the identification of unpopular users, especially those from the individual nodes, as reported in the study. In particular, the most useful metrics are those concerning

the concepts of centrality and popularity; respectively, they represent a measure of the importance of a node within a graph and the degree of trust which is attributed to it. The most effective metrics for this task are the following:

- Negated Number of Freaks (NNF): negated number of total foes of a node.
- Fans Minus Freaks (FMF): a user is called a "fan" of a friend and "freak" of an enemy. By subtracting the two values, it is possible to determine the reputation of an individual or a possible measure for the popularity.
- Page Rank (PR): measure that denotes the tendency of a person to be central, making no distinction between friends and enemies. It is therefore useful to define its popularity.
- Signed Spectral Ranking (SR): extended version of PageRank but aimed at measuring the popularity of a user on the network.
- Signed Symmetric Spectral Ranking (SSR): popularity measurement based on the idea that a popular user has few enemies and that the negative arches are more common among unpopular users.
- Negative Rank (NR): given the high correlation between the PR and SR measures, this additional metric is obtained as their difference.

To assess how the metrics are useful to identify unpopular users, researchers use them in order to flush out the trolls. Specifically, they create a dataset collecting the data regarding the enemies of a moderator account ("No More Trolls"). Evaluating the metrics for each of them, it is possible to notice that the measurement of NR gives the best results, followed by NNF.

Thanks to this new way of studying social networks and through other studies which explain how to transform any social network into one with "friends and enemies" [35], researchers have been able to proceed in this direction. As a result, several solutions to troll detection based on this approach were derived. For example, in [35] researchers try to improve this method by implementing an algorithm that, at each iteration, reduces the size of the social network. In particular, it iteratively deletes arches that are unnecessary for the analysis, focusing more on the types of "attacks" adopted by trolls. Instead, the work in [28] evaluates how it is possible to apply these studies on the propagation of trust, which is already covered in detail in the literature within the context of social networks, adapting them to those which admit arches with negative weights. The resulting approach allows for the diffusion of both confidence and distrust along the arches of the network. This way, it is possible to assign a measurement of "reliability" to each node, finally evaluating which users are trolls or not.

A potential service offered by the Twitter API is a timeline function, that returns the timeline of a given user, as its name suggests. The result includes the last twenty activities performed by the user, in JSON format. These are original tweets, retweets or quoted tweets. Specifically, this is a concatenation of more JSONs, each of them considered a single activity. It is worth noting that each JSON object provides a subsection that contains all information about the user's account, reporting many data about the owner of the activity. Each type of action (out of the three taken into consideration) is represented in a format that differs slightly from the others. In case of an original tweet, the JSON object has a simple structure, containing user information, text and tweet data. The retweet and quoted tweet actions have both a similar structure, because both actions are similar: the first action is a simple form of content sharing; the second action is mainly equal to the first, but it allows to add a comment to the message. In fact, both JSON objects have a part related to the owner user and a "sub JSON" part, referring to the original message, in turn containing the account information together with the text and their data; additionally, in the case of quoted tweets, the object contains the text comments along with their corresponding data.

The Twitter timeline API proves to be a great tool for obtaining information about a user, such as: total number of tweets, number of followers, number of friends, date of Twitter registration, total number of tweets, writing frequency, etc. Apart from this information, the timeline, being a report of the last twenty user's activities, can be viewed as a "snap-shot" of the actual state and used to detect the user behaviour, if appropriate metrics are considered. This statement is motivated also from the conclusions of Cheng et al. [37], their work says that a user in the process of being banned, during his

last period online, tends to write worse and drastically worsens his conflict with other people inside the community. This equates, basically, with a lower readability of his posts (data extracted from specialized metrics as ARI and Flesch-Kincaid [41]) and an exacerbation of his sentiment.

4. Conclusions

This article has discussed the problems of the presence of trolls in social media contexts and has presented the main approaches to their detection. The interest from the scientific community on the phenomenon of trolls and their automatic identification has emerged only recently, but the first research results show that the identification of malicious users in a social media context is possible. Different methods have been experimented with good results. However, different social media often offer different types and amounts of data which can be useful for the detection of trolls; therefore, such methods must be adapted and/or revised for their use in different kinds of community.

Our future work will start from the results of the works presented in this survey and from a tool for data analysis that we have recently developed and tested [42]. Our first experimentations were related to the Twitter communities and our future work will be dedicated to improving the solutions presented in [14] and to experiment the tool in other types of communities.

Author Contributions: Writing—original draft, M.M., M.T., and G.L.; writing—review and editing, S. C. and A. P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Larosiliere, G.; Carter, L.; Meske, C. How does the world connect? Exploring the global diffusion of social network sites. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 1875–1885.
2. Meske, C.; Stieglitz, S. Adoption and Use of Social Media in Small and Medium-sized Enterprises. In *Practice Driven Research on Enterprise Transformation (PRET), Proceedings of the 6th Working Conference Lecture Notes in Business Information Processing (LNBIP), Utrecht, The Netherlands, 6 June 2013*; Springer: Berlin, Heidelberg, 2013; pp. 61–75.
3. Meske, C.; Wilms, K.; Stieglitz, S. Enterprise Social Networks as Digital Infrastructures - Understanding the Utilitarian Value of Social Media at the Workplace. *Inf. Syst. Manag.* **2019**, *36*, 350–367.
4. Meske, C.; Junglas, I.; Stieglitz, S. Explaining the emergence of hedonic motivations in enterprise social networks and their impact on sustainable user engagement - A four-drive perspective. *J. Enterp. Inf. Manag.* **2019**, *32*, 436–456.
5. Chinnov, A.; Meske, C.; Kerschke, P.; Stieglitz, S.; Trautmann, H. An Overview of Topic Discovery in Twitter Communication through Social Media Analytics. In Proceedings of the 21st Americas Conference on Information Systems (AMCIS), Fajardo, Puerto Rico, 13–15 August 2015; pp. 4096–4105.
6. Stieglitz, S.; Meske, C.; Roß, B.; Mirbabaie, M. Going Back in Time to Predict the Future - The Complex Role of the Data Collection Period in Social Media Analytics. *Inf. Syst. Front.* **2018**, 1–15.
7. Meske, C.; Junglas, I.; Schneider, J.; Jakoonmäki, R. How Social is Your Social Network? Toward A Measurement Model. In Proceedings of the 40th International Conference on Information Systems, Munich, Germany, 15–18 December 2019; pp. 1–9.
8. Hardaker, C. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research Language Behaviour Culture* **2010**, *6*, 215–242.
9. Mihaylov, T.; Georgiev, G.; Nakov, P. Finding opinion manipulation trolls in news community forums. In Proceedings of the nineteenth conference on computational natural language learning, Beijing, China, 30–31 July 2015; pp. 310–314.
10. Badawy, A.; Addawood, A.; Lerman, K.; Ferrara, E. Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining.* **2018**, *9*, 31.
11. Badawy, A.; Lerman, K.; Ferrara, E. Who falls for online political manipulation? In proceedings of The Web Conference 2019 - Companion of the World Wide Web Conference, , San Francisco, CA, USA, 13–17 May 2019; pp. 162–168.

12. Chun, S.A.; Holowczak, R.; Dharan, K.N.; Wang, R.; Basu, S.; Geller, J. Detecting political bias trolls in Twitter data. In Proceedings of the 15th International Conference on Web Information Systems and Technologies, WEBIST 2019, Vienna, Austria, 18–20 September 2019; pp. 334–342.
13. Zannettou, S.; Sirivianos, M.; Caulfield, T.; Stringhini, G.; De Cristofaro, E.; Blackburn, J. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In Proceedings of the Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019, San Francisco, CA, USA, 13–17 May 2019; pp. 218–226.
14. Fornacciari, P.; Mordonini, M.; Poggi, A.; Sani, L.; Tomaiuolo, M. A holistic system for troll detection on Twitter. *Comput. Hum. Behav.* **2018**, *89*, 258–268.
15. Donath, J.S. Identity and deception in the virtual community. In *Communities in Cyberspace*; Routledge: Abingdon-on-Thames, UK, 2002; pp. 37–68.
16. Kirman, B.; Lineham, C.; Lawson, S. 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: New York, NY, USA, 2012; pp. 121–130.
17. Buckels, E.E.; Trapnell, P.D.; Paulhus, D.L. Trolls just want to have fun. *Personality and individual Differences* **2014**, *67*, 97–102.
18. Morrissey, L. Trolling is an art: Towards a schematic classification of intention in internet trolling. *Griffith Work. Pap. Pragmat. Intercult. Commun.* **2010**, *3*, 75–82.
19. Pfaffenberger, B. If I Want It, It's OK": Usenet and the (Outer) Limits of Free Speech. *The Information Society* **1996**, *12*, 365–386.
20. Herring, S.; Job-Sluder, K.; Scheckler, R.; Barab, S. Searching for safety online: Managing "trolling" in a feminist forum. *Inf. Soc.* **2002**, *18*, 371–384.
21. Galán-García, P.; Puerta, J.G.; Gómez, C.L.; Santos, I.; Bringas, P.G. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Log. J. IGPL* **2016**, *24*, 42–53.
22. Cambria, E.; Chandra, P.; Sharma, A.; Hussain, A. *Do not Feel the Trolls*; ISWC: Shanghai, China, 2010; Volume 664.
23. Derczynski, L.; Bontcheva, K. Pheme: Veracity in Digital Social Networks. In proceedings of the User Modelling and Personalisation (UMAP) Project Synergy workshop, CEUR Workshop Proceedings, Aalborg, Denmark, 7–11 July 2014; Volume 1181.
24. Dellarocas, C. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Manag. Sci.* **2006**, *52*, 1577–1593.
25. King, G.; Pan, J.; Roberts, M.E. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *Am. Political Sci. Rev.* **2017**, *111*, 484–501.
26. Luceri, L.; Giordano, S.; Ferrara, E. Don't Feed the Troll: Detecting Troll Behavior via Inverse Reinforcement Learning. arXiv **2001,** arXiv:2001.10570.
27. Romero-Rodríguez, L.M.; de-Casas-Moreno, P.; Torres-Toukoumidis, Á. Dimensions and indicators of the information quality in digital media. *Comunicar. Media Educ. Res. J.* **2016**, 24, 91–100. doi: 10.3916/C49-2016-09
28. Ortega, F.J.; Troyano, J.A.; Cruz, F.L.; Vallejo, C.G.; Enríquez, F. Propagation of trust and distrust for the detection of trolls in a social network. *Comput. Netw.* **2012**, *56*, 2884–2895.
29. Seah, C.W.; Chieu, H.L.; Chai, K.M.A.; Teow, L.N.; Yeong, L.W. Troll detection by domain-adapting sentiment analysis. In proceedings of the 2015 18th IEEE International Conference on Information Fusion, Washington, DC, USA, 6–9 July 2015; pp. 792–799.
30. Dollberg, S. The metadata troll detector, Swiss Federal Institute of Technology, Zurich, Distributed Computing Group, Computer Engineering and Networks Laboratory. Tech. Rep. Semester Thesis, 2015. Available online: https://pub.tik.ee.ethz.ch/students/2014-HS/SA-2014-32.pdf (accessed on 7 January 2020).
31. Younus, A.; Qureshi, M.A.; Saeed, M.; Touheed, N.; O'Riordan, C.; Pasi, G. Election trolling: Analysing sentiment in tweets during Pakistan elections 2013. In proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 411–412.
32. Hallman, J.; Lokk, A. Viability of sentiment analysis for troll detection on Twitter: A Comparative Study Between the Naive Bayes and Maximum Entropy Algorithms. KTH Royal Institute of Technology - School of Computer Science and Communication - Degree Project in Computing Engineering, Stockholm, Sweden. 2016. Available online: https://kth.diva-portal.org/smash/get/diva2:927326/FULLTEXT01.pdf (accessed on 7 January 2020).

33. de-la-Pena-Sordo, J.; Santos, I.; Pastor-López, I.; Bringas, P.G. Filtering Trolling Comments through Collective Classification. In *International Conference on Network and System Security*; Springer: Berlin, Heidelberg, 2013; pp. 707–713.
34. Bharati, P.; Lee, C.; Syed, R. Trolls and Social Movement Participation: An Empirical Investigation. 2018. Available online: https://pdfs.semanticscholar.org/fbd4/dc4eec69e6114cfd9011576f1f64c1bfbefc.pdf (accessed on 7 January 2020).
35. Kunegis, J.; Lommatzsch, A.; Bauckhage, C. The SlashDot zoo: Mining a social network with negative edges. In proceedings of the 18th International Conference on World Wide Web, city, country, day month year, 2009; pp. 741–750.
36. Dlala, I.O.; Attiaoui, D.; Martin, A.; Yaghlane, B. Trolls identification within an uncertain framework. In proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, Limassol, Cyprus, 10–12 November 2014; pp. 1011–1015.
37. Cheng, J.; Danescu-Niculescu-Mizil, C.; Leskovec, J. Antisocial behavior in online discussion communities. In proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015; pp. 61–70.
38. Kumar, S.; Spezzano, F.; Subrahmanian, V.S. (2014, August). Accurately detecting trolls in slashdot zoo via decluttering. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Beijing, China, 17–20 August 2014; pp. 188–195.
39. Atanasov, A.; Morales, G.D.F.; Nakov, P. Predicting the Role of Political Trolls in Social Media, 2019. Available online: https://arxiv.org/pdf/1910.02001.pdf (accessed on 7 January 2020).
40. Machová, K.; Kolesár, D. Recognition of Antisocial Behavior in Online Discussions. In *International Conference on Information Systems Architecture and Technology*; Springer: Cham, Switzerland, 2019; pp. 253–262.
41. Kincaid, J.P.; Fishburne, R.P., Jr.; Rogers, R.L.; Chissom, B.S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. University of Central Florida, 1975. Available online: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary (accessed on 7 January 2020).
42. Lombardo, G.; Fornacciari, P.; Mordonini, M.; Tomaiuolo, M.; Poggi, A. A Multi-Agent Architecture for Data Analysis. *Future Internet* **2019**, *11*, 49.