proceedings

Proceedings

Harassment Detection Using Machine Learning and Fuzzy Logic Techniques [†]

Jezabel Molina-Gil, José A. Concepción-Sánchez and Pino Caballero-Gil *

Department of Computer Engineering and Systems, University of La Laguna, 38271 La Laguna, Tenerife, Spain; jmmolina@ull.es (J.M.-G.); jaconcep@ull.es (J.A.C.-S.); pcaballe@ull.es (P.C.-G.)

* Correspondence: pcaballe@ull.edu.es; Tel.: +34-922-318176

† Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

‡ These authors contributed equally to this work.

Published: 20 November 2019

Abstract: Social networks, instant messaging applications, smartphones and the Internet are the main technological tools used by adolescents for communication. While they can benefit from those tools, they can also be used as a weapon for harassment. Cyberbullying is the name used for a current global social problem derived from harassment that uses offensive messages, which is severely affecting the youngest. Different types of software to identify and filter offensive contents have been developed in the last years. However, most of them are time consuming, not scalable and focused on very specific environments. To address this problem, we propose a mobile application for smartphones that provides a potential offensive content detection in order to determine whether a cyberbullying attack exists or not. In particular, we have developed an application that combines data pre-processing, fuzzy logic and machine learning to predict cyberbullying content. The main idea is to install a mobile application on the smartphone of a possible victim, so that it runs in the background. The system analyzes all received messages and notifications using data processing and decision-making algorithms. Finally, a fuzzy logic technique helps the system to reach a conclusion under a certain degree of imprecision.

Keywords: cyberbullying; data procession; fuzzy logic; machine learning; real time

1. Introduction

The use of smartphones grows year after year. This increase occurs in all ages, but younger participants, in addition, use their phones for longer. In fact, consumers aged 18 to 34 admit excessive use of smartphones. This use is mainly aimed at entertainment and social interactions, a fact that has helped create numerous problems among teenagers. Many of these problems still have no solution, and their consequences in many cases are devastating. We are talking specifically about cyberbullying, which unlike usual bullying where the harassment of the victim is carried out in person through physical or verbal attacks, involves bullies using means such as instant messaging applications, social networks and other methods through the Internet to harass the victim. Among the main consequences for the victim are emotional disorders, stress, depression, anxiety and even, in some cases, suicide.

Some data reveal that cyberbullying is a global problem. Teens claim to have been victims on social network of embarrassing or mean things posts about them. In the UK this problem has increased by 88% in recent years. Other countries, such as Brazil [1], registered that in 2016 cyberbullying had grown and 65% of those affected were women.

Adolescents are more likely to be negatively affected by biased and harmful contents than adults. For instance, in Argentina more and more young people aged 7-18 are affected by this problem. This is why many countries are now aware of its seriousness and have begun to take action. Hence, detecting online offensive contents to protect youngsters has become an urgent task. This is the case in Germany [2], which has created a bill that penalizes social networks that do not eliminate offensive content and humiliating messages. However, this is not enough because there are many ways to practice cyberbullying on the Internet. One of the most used is instant messaging applications, which are much more difficult to control. For this reason, there is an urgent need to find some solution that will make it possible to detect, eradicate and prevent this problem in the future.

This work presents a practical alternative that consists of a mobile application. It will be installed, in a hidden mode, on victims' smartphones. The application will be listening, in the background, at the received notifications. The content of these will be extracted and analysed to detect possible cyberbullying situations. The mobile application's decision making relies on three main techniques—data processing, fuzzy logic and supervised learning. First, data processing analyses the content searching for key words and discarding unnecessary ones. Second, fuzzy logic helps in decision making. It is a computational intelligence technique that allows working with information with a certain degree of imprecision. It works with intermediate values, simulating human thinking and are not based on a simple yes/no value. Third, supervised learning, which is one of the branches of machine learning, allows deduction of results using a previous training. Since it is not possible to use conventional logic to determine whether a potential victim is receiving cyberbullying, in this work, these techniques are used to determine the final result.

This paper is structured as follows. In Section 2, some works related to this proposal are mentioned. In Section 3, the operation and performance of the proposed mobile application are detailed. Section 4 describes the system based on data processing, fuzzy logic and machine learning for cyberbullying detection. Finally, the paper is closed with some brief conclusions and future work in Section 5.

2. Related Works

There are many works that seek for strategies to detect, prevent or combat cyberbullying. For example, in Reference [3] a qualitative study is presented that aims to explore, among a sample of students and school staff, the strategies used at an individual and general level to combat cyberbullying behaviour. Reference [4] includes some of the non-punitive approaches most commonly used in school and the workplace as well as some suggestions for working with parents. Finally, in Reference [5] the police are introduced as a complementary actor to parents, students, schools and services provided to combat this problem because, according to the study, police could help prevent cyberbullying by carrying out information tasks for students, parents and schools, creating information services, identifying perpetrators and helping victims.

Other authors have used text mining paradigms [6] on topics related to the detection of cyberbullying such as online sexual harassment [7] or vandalism [8]. However, there is not much research on cyberbullying detection. For example, Reference [9] describes a system for the detection and monitoring of cyberbullying cases from forums and communities. The system is based on the detection of three basic components of natural language—insults, swearing and informal addresses.

With regard to machine learning, papers such as Reference [10] can be found, where the authors present a methodology to detect and associate false profiles used for defamatory activities with a real profile within Twitter. As a result, they present a real success story for detecting cyberbullying in an elementary school. Reference [11] defends the need not to label young people as aggressors or victims of cyberbullying but rather to apply a degree of severity depending on their degree of harassment. They apply a classifier and then a fuzzy logic system that uses the classifier data to identify the severity of the harassment.

After searching related works, most proposals are focused on looking for patterns that can help detect cyberbullying. In addition, a few proposals for real applications to detect this problem are focused on very specific environments such as social networks like Twitter or Facebook. However, this is a problem because nowadays there are a lot of ways to cyberbully such as instant messaging applications or SMS in addition to social networks.

For this reason, the system proposed in this work does not focus on a specific environment but is intended to be used in the smartphones of potential victims, taking advantage of the push notifications they receive on their mobile phones. In addition, another advantage it presents is the use of both machine learning and fuzzy logic, which allows the system to rely on two different mechanisms to make the final decision whether a user is actually being victimized by cyberbullying.

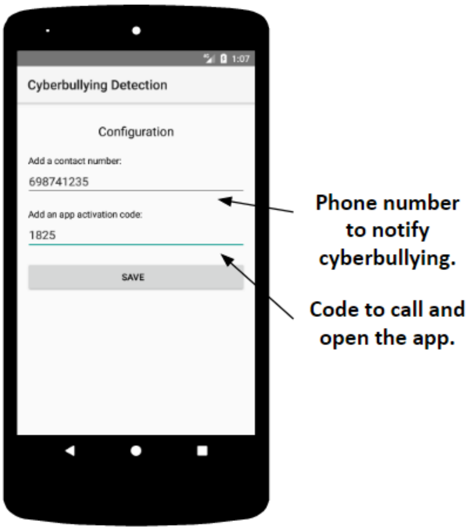
3. Proposed System

According to statistical data, it is expected that by 2020 there will be 6.1 trillion smartphones active in the world, which will be used by 70% of the global population [12]. In addition, the increased use of smartphones among young people [13] indicates that smartphones are becoming an essential part of society and that, in the worst case, they can be used for cybercrime. In this way, this work is focused on the development of an application for smartphones that makes use of a system for the detection of cyberbullying since it is the main means for doing this type of harassment.

The proposed application has been developed for Android devices, although the development for other mobile platforms is not ruled out. Its operation consists of the following steps:

First, since the application is intended to be used by parents or guardians who suspect that their children are victims of cyberbullying, they have to install the app on their children's smartphones. Note that the installation of the application is done in a hidden way (without icon) as this prevents the application from being detected by the harasser if he/she gets hold of the victim's smartphone.

Once the application is installed, parents have to add a call code to unlock the application and bring it to the foreground and a contact phone number (see Figure 1). Once this is done, when they click on the save button, the application will go into the background. Only if a call is made to the previous code, the application will unlock and return to the foreground. In this way, parents will only have to configure these parameters and wait to see if the application detects a possible cyberbullying case.



Add a contact number:
698741235

Add an app activation code:
1825

Phone number to notify cyberbullying.

Code to call and open the app.

Figure 1. Configuration screen of the mobile application.

Once the application is in the background, there is an associated listener that will allow detection of when a new notification has arrived to extract its content. Only if it detects a cyberbullying signal, it will notify the contact number previously configured via SMS so that appropriate measures can be taken.

Finally, since the data the application deals with are very sensitive because most of them are personal data and conversations, it is necessary to treat this information as carefully as possible so as not to invade the privacy of the user. In this way, no one, not even the parents themselves, will be able to obtain conversations from the children. Only in the case of detection of a cyberbullying case, the words or expressions that caused the alert will be sent via SMS to the contact telephone number.

That is why connections to the server for data analysis are made through HTTPS protocol. Besides, while there is not enough information to be analyzed, the application saves it as encrypted data in a database using the cryptographic algorithm AES 256 CBC mode [14], for later analysis.

4. System Operation

The system for detecting a possible cyberbullying case consists of the steps shown in Figure 2, where the used notation corresponds to:

- Data collection: It collects the data from the push notifications sent to the user for subsequent analysis.
- Data processing: It eliminates unnecessary words and returns the values needed by fuzzy logic and machine learning systems for analysis.
- Fuzzy logic system: Based on a set of linguistic input variables, it determines whether cyberbullying signs have been detected.
- Machine learning system: It deduces if the user could be a victim of cyberbullying from previous training.
- Final result: It consists of an AND operation between the results of the fuzzy logic and machine learning systems to determine if the user is finally being a victim of cyberbullying.

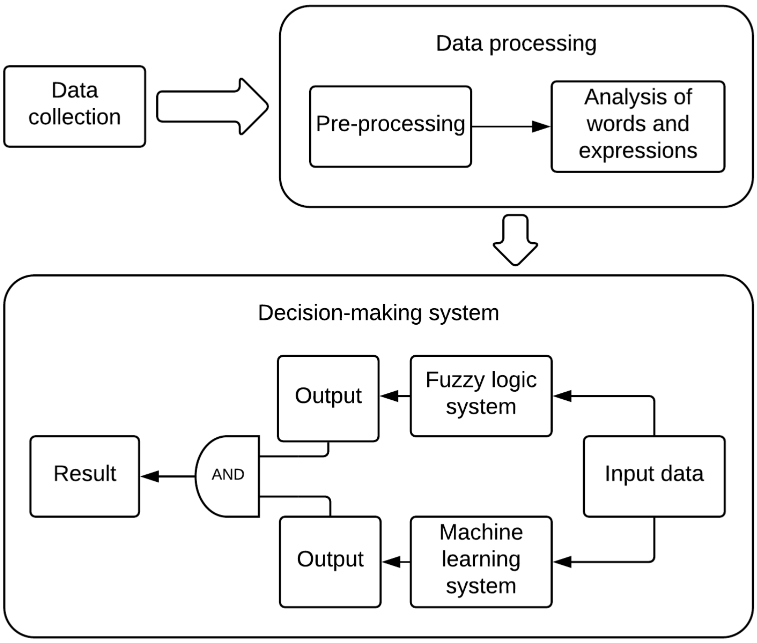
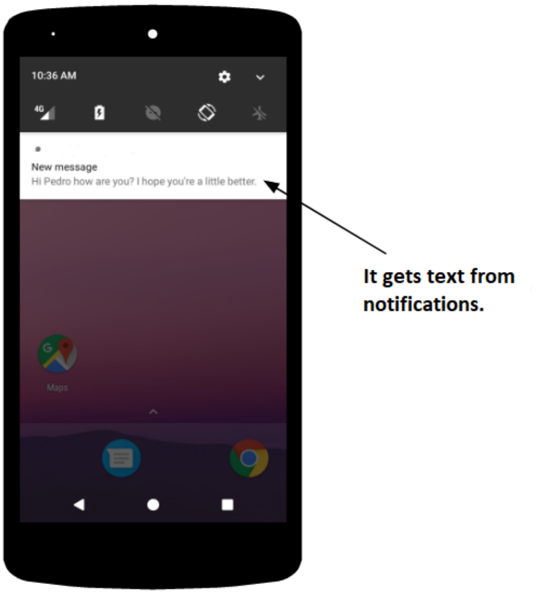


Figure 2. Diagram of the proposed cyberbullying detection system.

Each of the steps will be explained in more detail below.

4.1. Data Collection

To obtain the data, as was previously mentioned, the application parses the content of push notifications that reach the potential victim for further analysis. To do this, the application will be running in the background in a hidden way while waiting for new notifications (see Figure 3).



New message
Hi Pedro how are you? I hope you're a little better.

It gets text from notifications.

Figure 3. Mobile application running in the background.

Once the notification data has been collected, the application will check whether there is enough information to be analysed in the moment, proceeding with the next step of the proposed system. Otherwise, the application will store the encrypted data in a local database to analyze it when more information is available. This prevents the application from continuously analyzing the content, which could be harmful for the smartphone battery because sending and receiving data continuously from the server can consume a lot of power in the device.

Thanks to this mechanism, the application does not focus on a specific context but will be able to collect information from different sources that use push notifications. This could be the case of social networks such as Facebook or Twitter and instant messaging applications such as Telegram or WhatsApp.

4.2. Data Processing

Once the application has enough content stored for analysis in search of cyberbullying signals, the next step will be to process this information. In this step, unnecessary data will be removed and cyberbullying matches will be searched using a pre-fixed database of words and expressions that harassers could use.

4.2.1. Data Pre-Processing

The data pre-processing, done as described in Reference [15], is a very important step, as many times the useful content sent through social networks or instant messaging applications is negligible. In this way, pre-processing allows the elimination of unnecessary words such as articles, prepositions or other predefined words that do not have to be evaluated and classifying the remaining content into a set of words and expressions.

4.2.2. Data Analysis

To check the resulting words and expressions, a dictionary is used that contains possible words or expressions that a harasser could use to practice cyberbullying (see Figure 4). If no matches are found between the processed content and the dictionary, it would mean that there is no abusive content and therefore no need to continue with the next step as there would be no cyberbullying signs. Otherwise, the mobile application will make use of fuzzy logic and machine learning systems in the next step.

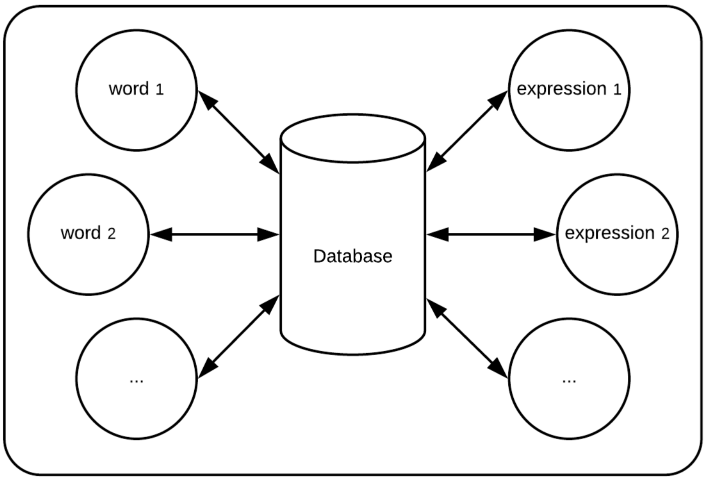


Figure 4. Data analysis searching matches.

4.3. Fuzzy Approach

Fuzzy logic is a methodology that provides a way to obtain a result from vague, ambiguous, inaccurate, noisy or incomplete input information. A fuzzy logic system is used because it allows a greater range of decision-making approaches where a deterministic system could not solve the problem. In this way, the system outputs are not only based on yes/no, but also considers a third variable we call incidence. This third variable is used when the system concludes that there is not enough information to generate a warning by cyberbullying but neither can the possibility be ruled out. In this way, the system will be able to take these incidences into account for future analysis and detect, for example, the repetition of a possible harassment for several days in a row or within a certain period of time.

The fuzzy logic system is composed of four linguistic input variables related to the text being analyzed:

- *Different Detections* (DD): Total number of different words and expressions that have been detected. The higher the number, the greater the likelihood that the user is being harassed.
- *Detection Frequency* (DF): Frequency of occurrence of detected words and expressions with respect to the total text. A higher frequent occurrence of words and expressions related to cyberbullying in the text may be a clear indicator that there is a case of cyberbullying.
- *Last Incidence* (LI): Days since the last time an incidence was generated. If the potential victim's application has not generated incidences in recent weeks, the likelihood of cyberbullying is reduced, while if the value is very small, the chances of cyberbullying increases.
- *Incidence Frequency* (IF): Frequency of incidence generation in the last three months. It is used to avoid isolated cases of incidences that could be false positives. In addition, a very high value of incidence frequency could be a symptom that the user is being harassed.

Each of these linguistic variables is composed of the linguistic terms HIGH, MEDIUM and LOW, which refer to the probability of cyberbullying detection according to each of the linguistic variables. Figure 5 shows the graphs where these linguistic terms are represented for each linguistic variable. The X coordinate corresponds to the values that can be taken by the linguistic variables and the Y coordinate with the probabilities corresponding to the linguistic terms. Each of these linguistic terms is represented by a membership function that define them. For our case, the triangular type membership functions have been used.

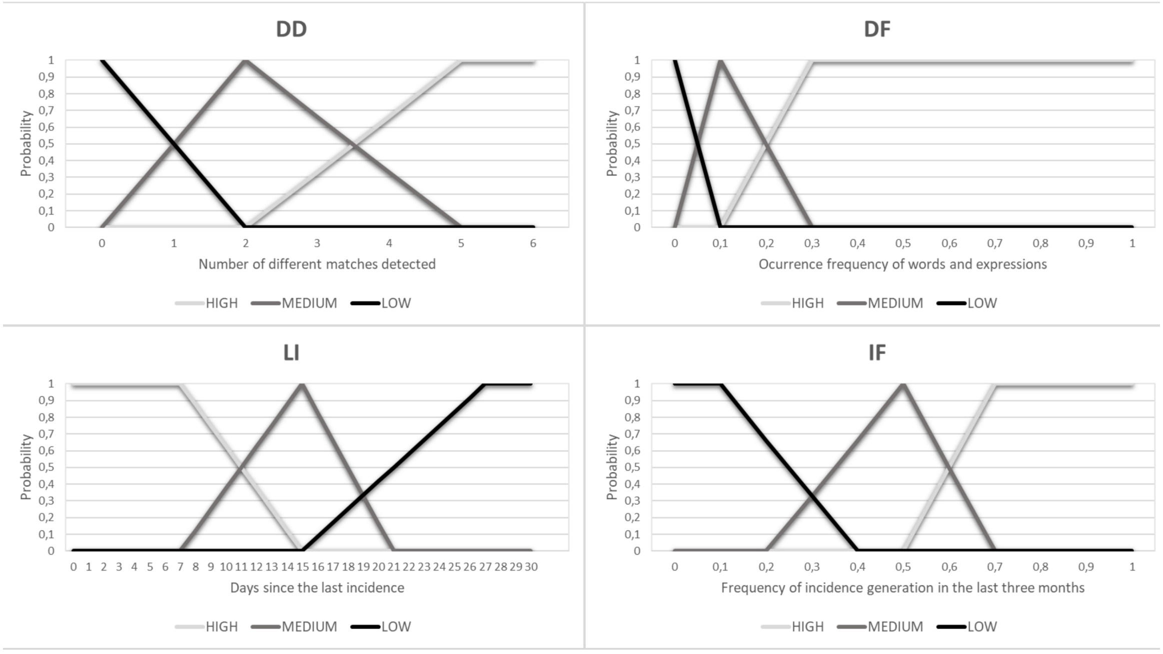


Figure 5. Graphs where the linguistic terms for each linguistic variable are represented.

The next step after fuzzification is to formulate specific rules for expressing the combination of influences. As an example, Listing 1 shows a very simple structure of rules where RESULT is the output language variable contained by three linguistic terms—YES, INCIDENCE and NO. These linguistic terms are associated with their corresponding membership functions, where the output will depend on which linguistic term has the maximum probability, using the max-membership defuzzification method. Besides, there may be more than one value assignment rule for the RESULT. In this case, the assignments to the RESULT are combined by an implicit AND, so the probability corresponding to the RESULT corresponds to the minimum value between all the input linguistic variable probabilities.

Listing 1: Sample rules of the fuzzy logic system.

```
INPUT: The fuzzified values of DD, DF, LI and IF.
OUTPUT: The fuzzified values of YES, INCIDENCE and NO.

IF (DD is HIGH) and (DF is HIGH) and (LI is LOW) and (IF is HIGH) THEN
     RESULT = YES;

IF (LI is HIGH) or ((DF is LOW) and (IF is HIGH)) THEN
     RESULT = INCIDENCE;

IF (IF is LOW) THEN
     RESULT = NO;
```

As an example, using the rules shown in Listing 1, if LI has four as value, its linguistic term HIGH will be fuzzified with 1. On the other hand, if IF has an incidence frequency of 0.6, its linguistic terms will be fuzzified with 0.5 as HIGH and 0.5 as MEDIUM. Once the linguistic terms are fuzzified, of the three rules established in the example, the first is the one that would be fulfilled, the reason why RESULT would have YES as a result. If more rules were fulfilled, they would be combined by AND, and the linguistic term with the highest probability will be chosen, as mentioned above.

4.4. Machine Learning System

Machine learning is a subfield of computer science and a branch of artificial intelligence whose objective is to develop techniques that allow computers to learn from data and then predict results based on them. Among the types of machine learning, supervised learning can be found, which allows deduction of a function from previous training data. These training data consist of pairs of objects, where one component is the input data and the other the desired result.

For this proposal, a previous training has been created using, as an input component, a set of data composed by variables that are also used as linguistic variables in the fuzzy logic system (DD, DF, LI and IF). For its part, a classification label is used to output the function, whose values are 0 (NO), 1 (INCIDENCE) and 2 (YES). Training data have been created manually by creating different scenarios of situations in which the texts and messages analysed have or do not have cyberbullying samples. In Table 1, an example of a training set table structure is shown:

Table 1. Sample training data.

DD	DF	LI	IF	Classification Label
1	0.001	14	0.05	NO
4	0.01	12	0.54	INCIDENCE
7	0.25	1	0.8	YES

Finally, the classification algorithm used to get the output from the input set is the decision tree algorithm. Among the advantages of this algorithm over other algorithms such as the Nearest Neighbor algorithm or the Naive Bayes classifier is the speed of prediction as well as the amount of resources needed to predict the result. Figure 6 shows an example of how the decision tree algorithm works.

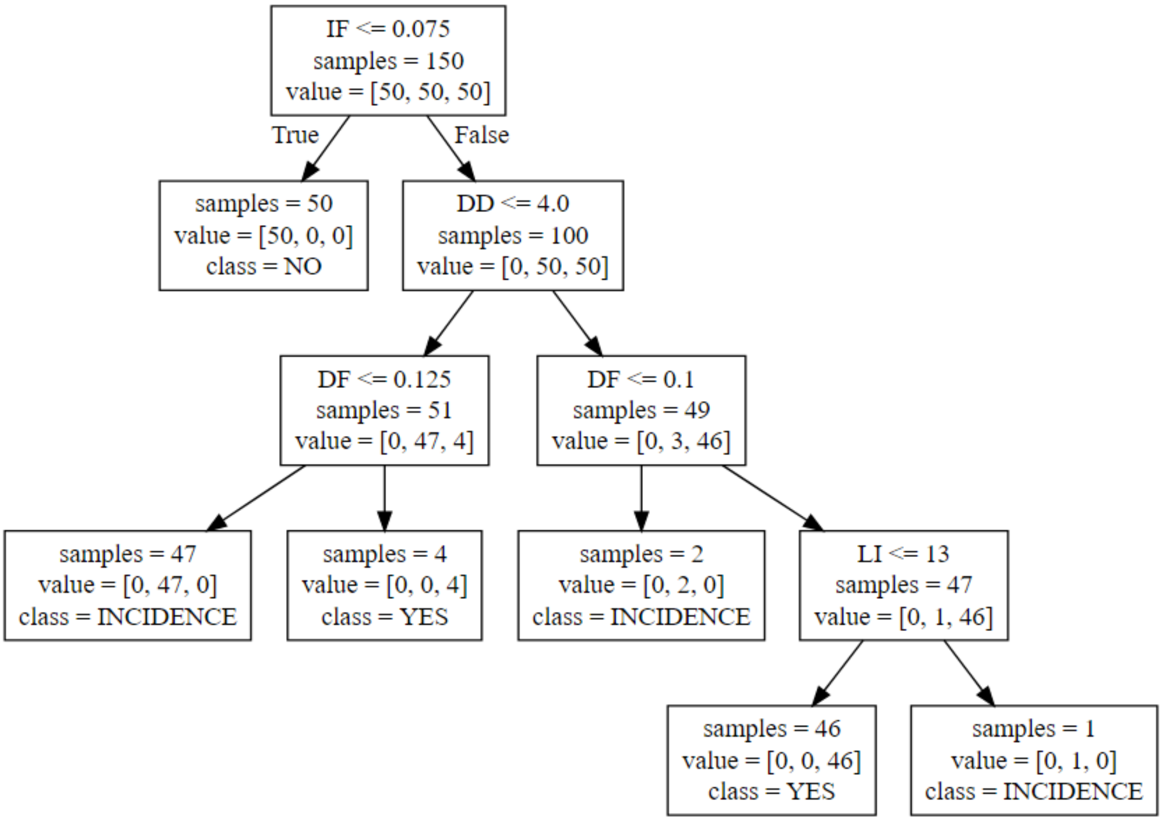


Figure 6. Sample decision tree algorithm.

4.5. Analysis of the Results

Due to the possibility that false positives may exist in some cases in the fuzzy logic or machine learning systems, in this last step an AND operation will be performed between the results of both systems so that Table 2) is applied on the AND of the results of both systems. In this way, only when the two systems return a new incidence or the detection of cyberbullying as a result, will the mobile application act accordingly. In all other cases, the lowest result will always be chosen between the two systems.

Table 2. AND operation between the results.

No	Incidence	Yes	Final Value
0	0	1	YES
0	1	0	INCIDENCE
0	1	1	INCIDENCE
1	0	0	NO
1	0	1	NO
1	1	0	NO

Finally, if the final result corresponds to the generation of an incidence, the application will take it into account for future analysis. On the other hand, if signs of cyberbullying are detected, parents will be notified along with the words and expressions that have been detected as mentioned above.

5. Conclusions and Future Works

In this paper we have discussed an important security problem that affects people and especially young people, which is cyberbullying through mobile phones. Most research on this topic has focused on finding patterns for the detection of cyberbullying through the effects it has on its victims. Besides, the few applications developed for this purpose focus on specific environments such as social networks.

The proposal presented in this work has provided a proof of concept to address this problem, consisting of an application that is installed on the potential victim's smartphone and looks to detect offensive content in received messages from all instant messaging applications. The application is installed in a hidden way and uses the push notifications that reach the smartphone to feed the system. The proposed implementation of the system consists of three steps. First, data processing is done by removing unnecessary tokens and searching for words or expressions with offensive content. Secondly, a fuzzy logic system has been implemented to improve decision making. The system is composed of four linguistic variables that, based on rules, allow conclusions to be reached based on inaccurate information. Finally, the system uses a machine learning mechanism to learn from data and predict results. Note that, although in this paper it has been assumed that the parents take care about the possible harassment of their children, it might be advisable for users (younger people) to take care of their own, and also in this case the proposed application could help them do this.

For future work, we envisage improving the functionality of the system by analysing another type of content that could be of great help. Specifically, one could add to the system the search for words that identify the user such as second person pronouns, or terms referring to people. This proposal arises because some studies show that when the offensive words are related grammatically to the identification of the users, the level of offense is greater. In addition, one could add criteria such as upper case and the use of exclamation marks because they are also used to increase the strength of the insult. On the other hand, with the aim of reducing false positives, a new variable to measure the offense intensity of a word could be added. This variable, with the two previous ones, could improve the fuzzy logic system results. Finally, the inclusion of new models, such as neural networks, will be studied so that the system can adapt to the natural language and vocabulary of each person and that in this way the false positives can be reduced even more.

Acknowledgments: Research supported by the Spanish Ministry of Economy and Competitiveness, the European FEDER Fund, and the CajaCanarias Foundation, under Projects TEC2014-54110-R, RTC-2014-1648-8, MTM2015-69138-REDT and DIG02-INSITU.

References

1. Itmidia. Vazamentos de Nudes Caem, mas Cyberbullying Cresce no Brasil em 2016. Available online: https://itmidia.com/vazamentos-de-nudes-caem-mas-ciberbullying-cresce-no-brasil-em-2016/ (accessed on 1 November 2019). 2017.
2. TheLocal.de. Germany to Fine Social Media Giants Up to €50 Million for Hate Speech. Available online: https://www.thelocal.de/20170405/germany-to-fine-social-media-giants-up-to-50-million-for-hate-speech (accessed on 1 November 2019). 2017.
3. Pelfrey, W.V., Jr.; Weber, N.L. Student and School Staff Strategies to Combat Cyberbullying in an Urban Student Population. *Prev. Sch. Fail.: Altern. Educ. Child. Youth* **2015**, *59*, 227–236, doi:10.1080/1045988X.2014.924087.
4. Bauman, S. Counseling Strategies to Combat Cyberbullying. In *Cyberbullying*; Wiley: Hoboken, NJ, USA, 2015; pp. 109–125. doi:10.1002/9781119221685.ch9
5. Vandebosch, H.; Beirens, L.; D'Haese, W.; Wegge, D.; Pabian, S. Police actions with regard to cyberbullying: The Belgian case. *Psicothema* **2012**, *24*, 646–652.
6. Chen, H.; Mckeever, S.; Delany, S.J. Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media. In *Advances in Computational Intelligence Systems*; Springer: Lancaster, UK; 2017; pp. 187–205. doi:10.1007/978-3-319-46562-3_12
7. Kontostathis, A. ChatCoder: Toward the tracking and categorization of internet predators. In Proceedings of the Text Mining Workshop 2009 Held in Conjunction with the Ninth Siam International Conference on Data Mining (SDM 2009), Sparks, NV, USA, 2009.
8. Smets, K.; Goethals, B.; Verdonk, B. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*; Chicago, USA; AAAI Press; 2008; pp. 43–48.
9. Foong, Y.J.; Oussalah, M. Cyberbullying System Detection and Analysis. In Proceedings of the IEEE Intelligence and Security Informatics Conference (EISIC), 2017 European, Athens, Greece, 11–13 September 2017; pp. 40–46.
10. Galán-García, P.; Puerta, J.G.d.l.; 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.
11. Sedano, C.R.; Ursini, E.L.; Martins, P.S. A Bullying-Severity Identifier Framework Based on Machine Learning and Fuzzy Logic. In *International Conference on Artificial Intelligence and Soft Computing*; Springer: Berlin/Heidelberg, Germany; 2017; pp. 315–324.
12. Mlot, S. *70 Percent of Population Will Have Smartphones by 2020*; PC Magazine; 2015. Available online: https://www.pcmag.com/news/334964/70-percent-of-population-will-have-smartphones-by-2020 (accessed on 1 November 2019).
13. Statista.com. Forecast of the Smartphone User Penetration Rate in the United Kingdom (UK) from 2015 to 2022. 2018. Available online: https://www.statista.com/statistics/553707/predicted-smartphone-user-penetration-rate-in-the-united-kingdom-uk/ (accessed on 1 November 2019).
14. Daemen, J.; Rijmen, V. *The Design of Rijndael: AES-the Advanced Encryption Standard*; Springer Science & Business Media; Springer: Berlin/Heidelberg, Germany; 2013.
15. Baskar, S.; Arockiam, L.; Charles, S. A systematic approach on data pre-processing in data mining. *Compusoft* **2013**, *2*, 335.