Textual Analysis for the Protection of Children and Teenagers in Social Media

Classification of Inappropriate Messages for Children and Teenagers

Thársis Salathiel de Souza Viana[1], Marcos de Oliveira[1], Ticiana Linhares Coelho da Silva[1],
Mário Sérgio Rodrigues Falcão Júnior[2] and Enyo José Tavares Gonçalves[1]
[1]*Universidade Federal do Ceará, Quixadá, Brazil*
[2]*Computer Science, Universidade Estadual do Ceará, Fortaleza, Brazil*

Keywords: Messages Exchange Classification, Social Media, Children and Teenagers Protection.

Abstract: Nowadays the Internet is widely used by children and teenagers, where privacy and exposure protection are often not prioritised. This can leave them exposed to paedophiles, who can use a simple chat to start a conversation, which may be the first step towards sexual abuse. In the paper (Falcão Jr. et al, 2016), the authors proposed a tool to detect possible dangerous conversations for a minor in a social network, based on the minor's behaviour. However, the proposed tool does not thoroughly address the analyses of the messages exchanged and attempts to detect the suspicious ones in a chat conversation using a superficial approach. This project aims to extend (Falcão Jr. et al, 2016) by automatically classifying the messages exchanged between a minor and an adult in a social network, hence to separate the ones that seem to come from a paedophile from those that seem to be a normal conversation. An experiment with a real conversation was done to test the effectiveness of the created model.

1 INTRODUCTION

Nowadays the Internet presents an imperative role in our society. From a simple source of entertainment to a fundamental tool behind a service or a company. Over time, social networks have stood out on the internet for allowing a new form of communication and social interaction. Some social networks incorporate instant messaging as an embedded feature. Thus, the social network user has not only a profile but also a means of communication with other users.

The messages exchanged between two users in a social network may occur in a public environment or a private one. Usually, conversations in a private environment tend to be about personal matters where the participating users do not aim at sharing those messages with other users. Children and teenagers are introduced earlier to this scenario and exhibited to private conversations with persons that they may not know. As a result, they can be more vulnerable to attacks from paedophiles or suspicious persons. Due to a large number of messages that are exchanged between users, it is impractical to check all messages manually. There is a need to classify those messages as dangerous or not automatically and inform if the child is the victim of a paedophile.

There exist papers that try to detect automatically messages that may come from paedophiles. The majority of them consider messages in English. The paper (Falcão Jr. et al, 2016) proposes a tool which analysis the profile behaviour of a minor and automatically identify the exposure level of a child based on his/her interactions. The tool analyses through heuristic information such as the number of dangerous friends, number of photos posted, groups, songs, period in which the user is logged in, among others. However, the textual analysis of the exchanged messages between the child and other users was addressed in a superficial manner by the approach of (Falcão Jr. et al, 2016). In this way, this paper aims to propose a textual analysis of the messages exchanged by children or teenagers of a social network, to detect if the profile examined was the victim of a sexual predator, and sum that analysis to the tool demonstrated in (Falcão Jr. et al, 2016).

Viana, T., Oliveira, M., Silva, T., Júnior, M. and Gonçalves, E.
Textual Analysis for the Protection of Children and Teenagers in Social Media - Classification of Inappropriate Messages for Children and Teenagers.
DOI: 10.5220/0006370606560662
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 2*, pages 656-662
ISBN: 978-989-758-248-6

The sections of this paper are organised as: Section 2, presents the related works, Section 3 presents our approach to solve the problem, Section 4 presents the results obtained, and finally Section 5 draws the conclusion and future work.

2 RELATED WORK

2.1 Behavioural Analysis for Child Protection in Social Network through Data Mining and Multiagent Systems

The software developed by (Falcão Jr. et al, 2016) uses a variety of information from the user's iteration, such as inadequate words, inappropriate books, videos or pages, etc. Assigning quantitative values for each information, proportional to the risk of exposure offered to the user. At the end, all generated values are summed up to calculate the level of exposure of the user. One of the pieces of information used to generate the classification model was the messages exchange with other users. However, this is done simply by looking for pejorative and sexual words or words that are not indicated for a child. The authors defined a kind of "blacklist" of words.

The first point where this check fails is in the case that a message may have some pejorative or sexual words which are not in the "blacklist". This fact is very likely to happen because some of these words are not even found in dictionaries, since in social network the users usually communicate by using informal words.

The second point that fails is to assume that conversations between the underage and the paedophile will always have words with pejorative or sexual content. The paedophile does not necessarily have to use those words to be able to act. He can pretend to be respectful until he achieves his goal, for example, to meet the child. A possible message received by an underage user may be: "Let's meet tonight." It is possible to notice that this message does not have any word with pejorative or sexual content, but it is a dangerous message if it comes from some adult. Thus, in many cases, the classification offered by the software fails, and it does not detect those messages considered dangerous for the child.

2.2 Identifying Online Sexual Predators by SVM Classification with Lexical and Behavioural Features

The paper proposed by (Morris, 2013) presents a similar approach to ours. The corpus used by them was assembled by the organisers of the PAN 2012 lab (Inches and Crestani, 2012). This corpus uses as a base for conversations considered dangerous, the website http://www.pervertedjustice.com, which is the same utilised by this work. The "normal" conversations are taken from other sites in English. Since it is a database of a conference, this database is larger than the one collected by this paper, but it is entirely in English. Since this article aims to classify texts in Brazilian Portuguese, the collection of "normal" conversations in Brazilian Portuguese was prioritised.

Some other differences are in the Lexical and Behavioural features used by (Morris, 2013) and in the preprocessing done by this paper. They make use of Behavioural features created from the metadata associated with the conversations. These features are divided into three types: Initiative (features that are related to the tendency to start the conversation with the partner), Attentiveness (features which are related to the author's willingness to keep the conversation) and Conversation dominance (features that are related as the author dominates the conversation). This paper, however, proposes an alternative text analysis to the software presented in (Falcão Jr. et al, 2016) that already performs a behavioural analysis based on several attributes. Another difference between the two papers is in the preprocessing phase. The paper of (Morris, 2013) uses lowercasing, stripping punctuation, and stemming for preprocessing. All of these routines are used by this paper as well. However, others are used to improve/clean the text to build a classification model. These routines are detailed in Section 3.2.

Finally (Morris, 2013) uses the algorithm SVM (support vector machine) while this paper uses the Multinomial Naive Bayes algorithm, which obtained a greater percentage of correctness.

3 OUR APPROACH

Our goal is to automatically detect if a message exchanged between a child and an adult is suspicious. To achieve our goal, in this paper we propose a model that analyses the messages of a minor's profile and classify it as dangerous or not. Our model is built from a classification algorithm which is used to predict one of the predefined categories (Tan, Steinbach and Kumar, 2005).

A classification algorithm for text is much more efficient than simply comparing if there is an inappropriate word in the message. Our proposal can identify patterns and classify more consistently. Also, by using a classification algorithm, it is possible to test it and measure the accuracy of the classification model, which can be useful to quantify the classification efficiency. Due to the large number of messages that can be exchanged between users, it is impractical to check all messages manually.

The chosen classification algorithm used in this paper is Multinomial Naive Bayes (Metsis, Androutsopoulos and Paliouras, 2006). The classification algorithms have been used by several papers to analyse text. The paper (Morris, 2013) performs the classification of messages which are particularly suggestive of predatory behaviour, using the SMV algorithm (support vector machine). The paper (Bogdanova, Rosso and Solorio, 2012) used the Naive Bayes to detect dangerous conversations in chats. The paper (McCallum and Kamal, 1998) makes a comparison between two probabilistic models for classification, multi-variate Bernoulli model and multinomial model, both of which make the Naive Bayes assumption. In our experiments, the Naive Bayes multinomial model was the one that obtained the highest percentage of correctness.

To run the algorithm, it is necessary to define previously in which categories the algorithm will classify the input (i.e., the messages). In this case, two categories were chosen: "normal" and "dangerous". In addition, it is also necessary to provide as input the training set and the test set. The following sections discuss these issues in more depth.

3.1 Data Collection

The dataset with the classified messages as "dangerous" was collected from the website: http://www.perverted-justice.com. This website has several exchanged messages between a paedophile and an activist who pretends to be an underage user in a public chat. The dataset with the messages classified as "normal" was retrieved from the website: http://vircio.net. This website has many conversations records on several topics.

In the experiments, we have used 1610 messages in total ("dangerous" and "normal" messages). Since in our dataset, the sentences considered dangerous generally contained more words, the number of messages classified as "dangerous" and "normal" was not exactly the same. To be fair, we took into consideration the number of words and characters in the sentences. In this way, 702 messages were classified as "dangerous" (4852 words, 24350 characters) and 908 messages as "normal" (4598 words, 24055 characters).

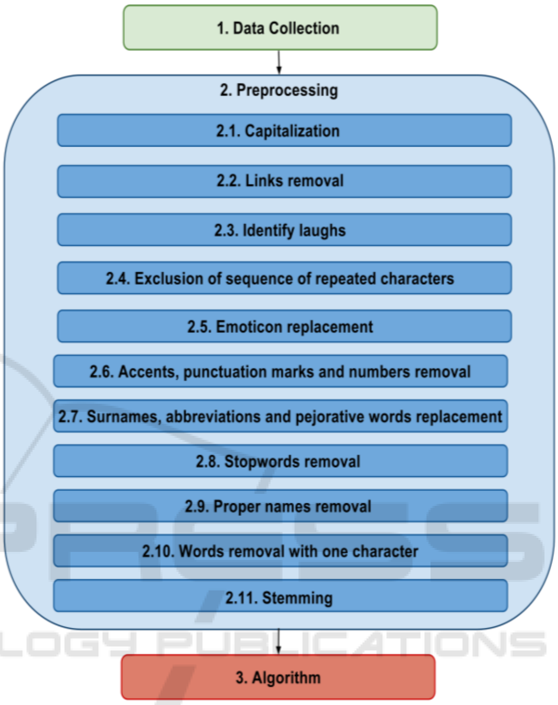

1. Data Collection
2. Preprocessing
2.1. Capitalization
2.2. Links removal
2.3. Identify laughs
2.4. Exclusion of sequence of repeated characters
2.5. Emoticon replacement
2.6. Accents, punctuation marks and numbers removal
2.7. Surnames, abbreviations and pejorative words replacement
2.8. Stopwords removal
2.9. Proper names removal
2.10. Words removal with one character
2.11. Stemming
3. Algorithm

Figure 1: Steps of our approach.

The messages were split into the training and test set, where 30% of the dataset was for test and 70% for training. We experimented with the k-fold cross-validation method, which randomly divides the data into k equal sets. Then, it chooses one of these "k" sets to be the test set and the other parts compose the training set. This processing is repeated k times. In order to obtain the classification statistics, the algorithm computes the average considering all the results.

3.2 Preprocessing

For the classification, the data used as training set were initially in English, so the messages were translated to Portuguese as a preprocessing step. Moreover, we also applied the Natural Language Processing as we discuss in the following paragraphs.

The messages collected have characteristics that differ from formal conversations. Grammar rules and punctuations are rarely used, words are often spelt wrong, and there is a frequent use of slang and abbreviations. This lack of pattern makes preprocessing more difficult since the same word can appear in several ways. For example, the user can repeat some specific letter to imitate a spoken language as in "hiii" (the correct word is "hi"). Also, the words in this kind of scenario might be typing with an accent missed, by mistake of the user or with the purpose to save time. Another possible scenario is one in which the user uses an abbreviation instead of the word itself.

There are other problems, such as punctuation. Punctuation as a semicolon, exclamation, interrogation, points and commas are hardly used. Even when they are used, there is no guarantee that they are used correctly. Another problem is that these characters also appear in the use of emoticons, which are graphical representations of feelings, such as: ":-)" (happy), ": @" (angry) and ":-(" (unhappy).

All these particularities are challenges for the classification processing. The solution adopted in this paper was to identify the main problems and to create a sequence of methods to standardise the sentences. Some of these methods were proposed in related works and others were created by the authors of this work. Methods such as lowercasing, stripping punctuation, and stemming were also used by (Morris, 2013). They have proposed replacing emoticons and proper names, although this paper differs from the word chosen for the replacement. The methods to remove the accents and the stopwords were also used by (Leite, 2015).

The list of preprocessing techniques applied, as described in Figure 1:

1. Capitalization: every character in lowercase.
2. Links removal.
3. Identify laughs and replace with a single word that represents laughter. The first part of this step is to identify a laugh. In Brazilian Portuguese, a laugh has more of a way of being written like: "hahaha", "kkkkk", "huehue". After identifying some of these several words that represent laughter, it is replaced by a unique word that will represent laughter. So different forms of laughter are transformed into a simple form. Thus the algorithm can more easily identify a word that represents a laugh.
4. Exclusion of sequence of repeated characters. The user can miss typing and put several characters in sequence. Sometimes he/she does it on purpose to express a way of speaking. In both cases, the repeated characters are transformed into one.
5. Emoticon replacement by a word that represents that feeling. It is easier for the classifier to work only with words. For example: ":-)" is replaced by "happy".
6. Accents, punctuation marks and numbers removal.
7. Surnames, abbreviations and pejorative words replacement by an appropriate word. Some nicknames and abbreviations are well known and widely used in chats, so these nicknames and abbreviations have been replaced by the word they represent. In this way either write abbreviations or the correct word, it will represent a unique word at the end of this step. Some pejorative words with several variations were also replaced by a unique word.
8. Stopwords removal. Stopwords are words that are not relevant for the classification processing, like articles and prepositions.
9. Proper names removal. Proper names are also not relevant for this analysis.
10. Words removal with one character. For some typing error, the user may have typed an isolated character, this will have no relevance in the classification.
11. Stemming. In Brazilian Portuguese some suffixes are added at the end of a word to generate another word, called "derived word". For example, consider the word "Cachorro" (dog in Portuguese). We can also have the words "Cachorrinho" (small dog in Portuguese), "Cachorrão" (big dog in Portuguese), "Cachorra" (female dog). All these words have essentially the same meaning. In the classification, we are only interested in the "root" of the word. The words "Cachorro", "Cachorrinho", "Cachorrão", "Cachorra", will be transformed into "Cachorr" ("root") at the end of this step.

3.3 The Algorithm

The Naive Bayes algorithm is well known in machine learning. The algorithm uses Bayes' theorem to calculate the probability of an attribute belonging to a particular class. It is called a "naive" because it assumes that the attributes are independent, a naive premise. In the text document classification each attribute to be classified are the words in the document.

The Naive Bayes classifier has two different models. The binary model where the document is represented by a vector of binary attributes indicating which words occur and do not occur in the document. The number of times the word appears in the document is ignored.

And the multinomial model where a document is represented by the set of word occurrences from the document, considering the number of occurrences of each word in the document.

The multinomial model seems to be more realistic for this problem. Because a word considered "dangerous" may appear in a "normal" conversation sometimes, but it will be more frequent in dangerous conversations. For example, the word "kiss". It is a word that appears a lot in paedophile conversations, but it can appear as well in a normal conversation between friends or parents and their children.

4 CHALLENGES

The main challenges faced in the development of this work were found in the data collection stage. A challenge already mentioned was the translation into Portuguese of the messages from the site http://www.perverted-justice.com. Unfortunately, there is no open database in Brazilian Portuguese for the collection of paedophile messages, so the translation phase is necessary. The messages often contained slang and abbreviations in English, then the automatic translation processing is not convenient.

From the authors knowledge, there is no dataset available with "typical" conversation between a minor and an adult. That is another challenge that we faced in this work. The solution found by the authors was to gather messages on several topics, where adults usually participate. All these messages which were used to create the model were manually checked to avoid adding a malicious message and classify it as normal.

5 RESULTS

To train, test the model and generate its statistics, the WEKA tool was used. These were the results for the model achieved by using Naive Bayes algorithm:

Table 1: Results of the trained model.

Total Number of Instances	1610
Correctly Classified Instances	1390 (86.3354%)
Incorrectly Classified Instances	220 (13.6646%)
Kappa statistic	0.7238

The first thing to note is the high percentage of instances ranked correctly, approximately 86%. The second point to note is the Kappa coefficient that means it can be considered a suitable model (between 0.60 and 0.80, according to (Carletta, 1996)).

The resulting confusion matrix was:

Table 2: Confusion matrix.

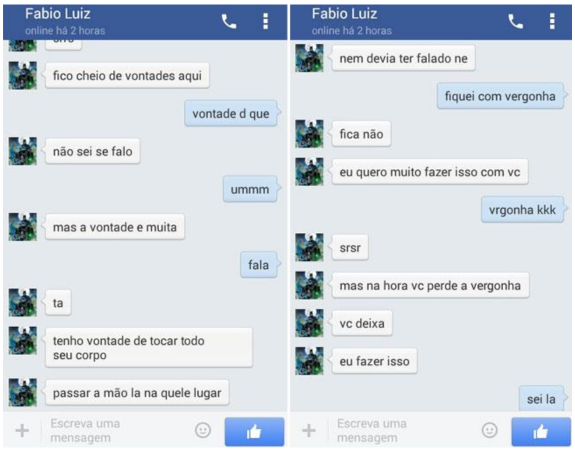
	dangerous (predicted)	normal (predicted)
dangerous (actual)	610	92
normal (actual)	128	780

When we analyse the confusion matrix we can see that the smallest number is the number of dangerous instances classified as normal (we call the false positives). For a classification of dangerous messages, it is important that the number of false positives be lower, since it means the number of messages that are dangerous but they will be classified as normal by the model. Thus making it impossible for a possible paedophile to be uncovered.

A slightly larger number is that of normal messages that are classified as dangerous (we call. the false negatives). This number does not concern as the false positives because the classification of a message as dangerous will only result in the isolation and registration of this message. Hence, it can be analysed after by a specialist who can recognise if the message really belongs to a paedophile.

5.1 A Real Case

Unfortunately, it is not easy to get access to real paedophile conversations with children, because the police generally do not make these conversations available. But some cases at least parts of the conversations are published. So we used this small conversation dataset in order to test the effectiveness of our proposed classification model with a real case. We conducted an experiment with part of a real conversation between a 35-year-old paedophile and a 9-year-old girl. The case happened in the city of Juiz de Fora, in the state of Minas Gerais, the suspect was arrested. A chat screenshot reveals a part of the exchange of messages between the paedophile and the child.



fico cheio de vontades aqui

vontade d que

não sei se falo

ummm

mas a vontade e muita

fala

ta

tenho vontade de tocar todo seu corpo

passar a mão la na quele lugar

nem devia ter falado ne

fiquei com vergonha

fica não

eu quero muito fazer isso com vc

vrgonha kkk

srsr

mas na hora vc perde a vergonha

vc deixa

eu fazer isso

sei la

Figure 2: Conversation between paedophile and 9-year-old child. Retrieved October 2, 2016 from http://www.tribunademinas.com.br/suspeito-de-pedofilia-flagrado-por-pai-de-vitima-no-facebook-esta-no-ceresp/.

These messages were classified by the model, which obtained the following results:

Table 3: Messages from the paedophile, the translation and the result of the classification.

Message	Translation	Result
fico cheio de vontades aqui	I'm full of desires here	Normal
não sei se falo	I do not know if I say	Normal
mas a vontade e muita	But the desire is too much	Normal
tenho vontade de tocar todo seu corpo	I want to touch your whole body	Dangerous
passa a mão la na quele lugar	Put my hand in that place	Dangerous
nem devia ter falado ne	I should not have even talked	Normal
fica não	Do not stay	Normal
eu quero muito fazer isso com vc	I really want to do this with you	Dangerous
mas na hora vc perde a vergonha	But at the time you lose the shame	Normal
vc deixa eu fazer isso	You let me do this?	Dangerous

By analysing the messages classified as normal, without analysing the context, it is difficult even for a person to decide whether they come from a paedophile or not. The only message classified as normal that seems to have come from a paedophile is the message: "mas na hora vc perde a vergonha" (translation: "But at the time you lose the shame"), in this case apparently there was an error on the part of the classifier. However, all messages classified as dangerous actually appear to have come from a paedophile.

This example shows something very important. Usually, a paedophile sends not just a single suspicious message, but several. The probability of a model classify all dangerous messages in a conversation as normal is very low. All in all, it is enough to have a single message classified as dangerous to be isolated and registered, and then the user that sent it be considered suspicious.

6 CONCLUSIONS AND FUTURE WORKS

Compared with the old system, it simply checked if the text had words from a list of "forbidden" words. The new way of classifying is much more robust and closer to reality.

This paper is also significant because it presents a tool to identify dangerous messages in Brazilian Portuguese, as previously said other works have similarity to this, but the vast majority propose models addressed for conversations in English. It is important that a country like Brazil with a large population has tools that help protect children and teenagers.

The goal is to gather this work to the work of (Falcão Jr. et al, 2016). However, the work of (Falcão Jr. et al, 2016) is focused on the social network Facebook, which disabled access to inbox messages via API. Nevertheless, this approach can make use of other ways to collect the data and tests will be applied in other social networks, like Twitter and Google+. Data can also be obtained by manually collecting messages from children's profiles on Facebook, with the permission of their parents.

As soon as we gather more data and rebuild the classification model, the accuracy of the model should be improved and covers more cases or subjects. In this case, it will be possible to keep the dataset modernised according to the varied linguistic strategies used by children sexual abusers.

REFERENCES

Falcão Jr., M. S. R., Gonçalves, E. J. T., Silva, T. L. C. da S., de Oliveira, M., 2016. Behavioral Analysis for Child Protection in Social Network through Data

Mining and Multiagent Systems. *In Proceedings of 18th International Conference on Enterprise Information Systems (ICEIS 2016)*. SCITEPRESS.

Inches, G., Crestani, F., 2012. Overview of the International Sexual Predator Identification Competition at PAN - 2012. In *CLEF (Online working notes/labs/workshop)*. Vol. 30.

Morris, C., 2013. Identifying online sexual predators by svm classification with lexical and behavioral features. *Master of Science Thesis*, University Of Toronto, Canada.

Leite, J. L. A., 2015. Mineração de textos do twitter utilizando técnicas de classificação. *Monografia de Final de Curso*. Univercidade Federal do Ceará, Campus Quixadá.

Tan, P., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining. Addison-Wesley, 1° Edition.

McCallum, A., Kamal, N., 1998. A comparison of event models for naive bayes text classification. *In Proceedings of AAAI-98 workshop on learning for text categorization*. Vol. 752.

Bogdanova, D., Rosso, P., Solorio, T., 2012. On the impact of sentiment and emotion based features in detecting online sexual predators. *In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics.

Metsis., Androutsopoulos, I., Paliouras, G., 2006. Spam filtering with naive bayes-which naive bayes?. CEAS. Vol. 17.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Computational linguistics Vol. 22.2. pp 249-254.