Automatic Detection of Insulting Sentences in Conversation

Merav Allouch
Computer Science Department
Ariel University
Ariel, Israel
merav@g.jct.ac.il

Amos Azaria
Computer Science Department
Ariel University
Ariel, Israel
amos.azaria@gmail.com

Rina Azoulay
Dept. of Computer Science
Jerusalem College of Technology
Jerusalem, Israel
rrinaa@gmail.com

Ester Ben-Izchak
The Department of Communication Disorders
Ariel University, Ariel, Israel
benitze@ariel.ac.il

Moti Zwilling
Department of Economics and Business Management
Ariel University, Ariel, Israel
motiz@ariel.ac.il

Ditza A. Zachor
Department of Pediatrics, Assaf Harofeh Medical Center
Sackler Faculty of Medicine Tel Aviv University, Israel
dzachor@bezeqint.net

***Abstract*—An overall goal of our work is to use machine-learning based solutions to assist children with communication difficulties in their communication task. In this paper, we concentrate on the problem of recognizing insulting sentences the child says, or insulting sentences that are told to him. An automated agent that is able to recognize such sentences can alert the child in real time situations, and can suggest how to respond to the resulting social situation.**

We composed a dataset of 1241 non-insulting and 1255 insulting sentences. We trained different machine learning methods on 90% randomly chosen sentences from the dataset and tested it on the remaining. We used the following machine learning methods: Multi-Layer Neural Network, SVM, Naive Bayes, Decision Tree, and Tree Bagger for the task. We found that the best predictors of the insulting sentences, were the SVM method, with 80% recall and over 75%precision, and the Multi-Layer Neural Network and the Tree Bagger, with precision and recall exceeding 75%, We also found that adding additional data to the learning process, such as 9500 labeled sentences from twitter, or adding the word "positive" and the word "negative" to sentences including positive or negative words, respectively, slightly improves the results in most of the cases.

Our results provide the cornerstones for an automated system that would enable on-line assistance and consultation for children with communication disabilities, and also for other persons with communication problems, in a way that will enable them to function better in society through this assistance.

Keywords Autism Spectrum Disorder, Machine Learning, Text Emotion Recognition

I. Introduction

Autism spectrum disorder (ASD) is a lifelong neuro-developmental disorder characterized by impaired reciprocal social communication and a pattern of restricted, often non-adaptive repetitive behaviors, interests and activities [2].

One of the widely accepted cognitive explanations for these symptoms is deficits in theory of mind (ToM) in ASD. ToM refers to the ability of individuals to impute mental states such as emotions, beliefs and ideas to oneself and to others and to predict the behavior of others on the basis of their mental states [3], [16]. ToM performance is a crucial capacity which enables one to decode and understand social cues [9]. Difficulties in ToM performance can impair social interactions including deficits in pragmatic abilities and empathy [13]. These deficits might lead one to declare innocently insulting statements or to misperceive bullying expressions directed to himself. High prevalence of bullying toward children and adults, including verbal bullying such as name calling, teasing and others, was documented in ASD (summarized in [14]).

978-1-5386-5541-2/18/$31.00 ©2018 IEEE

Children with special needs, and, in particular, children with autism spectrum disorder (ASD), may find difficulties in their interactions with other people, and in the understanding of social situations. Their disorder challenges their ability to interact with family members, peers, and teachers. They also find it difficult to identify social situations, feelings, expressions, etc. These functional difficulties increase the risk of social exclusion, where children with ASD may experience rejection, bullying and isolation [4].

In our research, we intend to concentrate on the task of developing an agents-based system to assist children with special needs in their communication with other people. In order to help these children, an automated agent will be aware of the child's interactions, and will give him relevant feedback and suggest appropriate responses to several possible situations.

A special situation that can challenge the child with ASD is a situation in which the child says an insulting sentence unintentionally. Such situations are, unfortunately, very common among ASD children. In particular, parents often report that their child can express themselves with sentences like "you are fat", "you are old", "go home", "the food stinks" without realizing that they are insulting. In this study, we suggest a method to recognize insulting sentences that were told by the child, or insulting sentences told to him by other children or other people. Thus, the aim of our research is to design an automated agent that will be able to detect insulting sentences, in order to be able to provide the child relevant feedback when such insulting sentence or sentences are detected.

We composed a dataset of insulting and non-insulting sentences with 1241 non-insulting and 1255 insulting sentences. The dataset was composed using the following method. An initial seed of 100 unintentional insulting sentences was obtained by performing interviews with parents of children with ASD (performed by the Autism Center). To this seeding dataset, and we added both insulting and non-insulting sentences from varied sources, including forums and article comments, with focus on sentences that can be said by children, or to a child.

We divided the sentences of the datasets into three parts: sentences that are certainly insulting, sentences that are not insulting at all, and sentences that may be insulting depending on the situation and context, including the place and time where and when the sentence was said, and who heard it. In this study we focus on determining insulting or non-insulting sentences, while we leave the goal of detecting the "Context-sensitive sentences" for future work.

We divided the sentences set to 90% training set and 10% test set, using a random partition. Then, we checked the abilities of different machine learning methods to predict the insulting sentences of the test set. We found that the best predictors for the insulting sentences, were the SVM method, with 80% recall and precision up to 75%, and the Multi-Layer Neural Network and the Tree Bagger, with precision and recall exceeding 75%, namely more than 75% of the insulting sentences can be recognized, and in addition, more than 75% of the sentences that are detected as insulting are indeed insulting sentences.

The rest of the paper is organized as follows. Section II describes some related work regarding current methods used for sensitive analysis and hate-sentence detection. Section III describes the dataset we developed, the machine learning methods we used, and the accuracy results for different models. Section IV concludes and suggests directions for future research.

II. Related Work

Children and adults with ASD may encounter difficulties in their communication with other people and in understanding the social situations. In recent years, several Information Communication Technology-based methods were developed and used for the therapy and education of children with ASD. Boucenna et al. [5] provide a comprehensive review on technologies, algorithms, interfaces and sensors that are able to sense the behavior of the children, train and improve their social abilities and train individuals to recognize facial emotions, emotional gestures and emotional situations. Boucenna et al. suggest the use of robots to provide feedback and encouragement during skill learning intervention; they emphasize that a child with ASD might find it easier to interact with a robot than with a human teacher. They suggest that the robot provide instructions to the child to interact with a human therapist and encourage the child to proceed with the interaction. Boucenna et al.'s research was theoretical, and in our research, we intend to develop such an assistant artificial agent, where the first step is to have the agent understand the current social situation, given the last sentence/s that was/were said by him or to him.

In order to implement the idea of an automated assistant, we need to solve the relevant algorithmic challenges. First of all, the automated assistant should be able to recognize problematic situations that the child encounters, and, in particular, recognize insulting sentences said by the child with ASD, or said to him.

Thus, we proceed in this overview with some related work on text classifying methods, and we concentrate on work on emotion recognition and sarcasm recognition.

Sentiment analysis is the task of determining the sentiments and emotions of the writer or a speaker of some text or speech [12]. Abbasi et al. [1] use sentiment analysis methodologies for classification of Web forum opinions in multiple languages. For this goal, they consider a wide array of stylistic attributes, including lexical, structural, and word style markers, in addition to syntactic features, and they used a hybridized genetic algorithm that incorporates the information-gain heuristic for feature selection. Gao et al. [10] use a rule-based system underlying the conditions that trigger emotions based on an emotional model . The data set comprised Chinese micro-blogs. They used the ECOCC emotion model and extracted the corresponding cause components in fine-grained emotions. An F-score of 0.7 and more was achieved. Nobada et al. [15] use a Vowpal Wabbits regression model and NLP features to detect hate speech on online user comments from two domains which outperforms a state-of the-art deep learning approach. Their features are divided into four classes: N-grams,Linguistic, Syntactic and Distributional Semantics. The dataset comprised Yahoo data. Kharde et al. [11] survey work on sentiment analysis, and concentrate on twitter data in which it is harder to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are positive or negative, and in some cases neutral. They provide a comparative analyses of existing techniques for opinion mining, like machine learning and lexicon-based approaches, together with evaluation metrics. Gui et al. [7] introduce the issue of emotion cause extraction, which means extracting the stimuli, or the cause of an emotion. Due to the lack of open resources for this area of study, they first constructed an annotated data set based on 3 years (2013-15) of Chinese city news. Then, they proposed an event-driven emotion cause extraction method to capture the emotion cause extraction. They proposed a 7-tuple representation of events using syntactic structures to identify events. Based on this structured representation of events and the inclusion of lexical features, they designed a convolution kernel based learning method in order to identify the emotion cause events. Our work is different from sentiment analysis because in sentiment analysis the emotions of the writer are detected, and in our work, we focus on detecting the sentences that cause the listener to get insulted. It is also different than hate-speech detection, because the insulting sentences in our domain can be the result of innocent intentions, and in most cases the do not contain sentences that are considered hate speech.

III. Insulting Detection: System Description and Experimental Results

As described above, the first step in assisting the child in his communication task is to help him detect insulting sentences spoken to him or spoken by him. Available on-line data set resources of essays, comments and recommendation are not entirely appropriate for our goal, since the terms used in on-line spoken conversation at home, class, near friends, etc., can be different from the terms used in written text of comments, etc., especially when considering conversation of children. In addition, on-line resources can help us identify hate terms or to extract emotions from text, but it is more difficult to find on-line insulting text when observing on-line texts or essays which are intended for the public and not a single listener or a defined group of listeners.

Thus, we first designed a dataset of sentences which contains 1241 insulting sentences and 1255 non-insulting sentences. The dataset was composed using the following method. An initial seed of 100 unintentional insulting sentences was obtained by performing interviews with parents of children with ASD (performed by the Autism Center). To this seeding dataset, we added both insulting and non-insulting sentences from varied sources, including news, forms, descriptions of situations written by participants of several on-line support groups, etc. with focus on sentences that can be said by children, or to a child.

We labeled the data according to three types of sentences: clearly insulting sentences, clearly non-insulting sentences, and sentences which may be insulting or non-insulting, depending on the situation. (For example, a sentence like "When do you clean your room?", may depend on the context where and when it is said). In this study we concentrate on detecting clearly insulting and clearly non-insulting sentences, and we leave the task of detecting context-sensitive sentences to future research.

Given the insulting/non-insulting dataset, we posed the following questions:

1) Can a learning method trained by an available Twitter hate-speech dataset[1] successfully predict insulting sentences in our dataset?
2) Can a learning method trained by part of the insulting/non-insulting dataset predict the insulting

[1]https://data.world/thomasrdavidson/hate-speech-and-offensive-language

sentences in reference to the rest of the dataset, and will it improve its predicting ability when trained by both the Twitter hate-speech dataset and part of the insulting/non-insulting dataset?

3) Which learning method is the best predictor?
4) Can we find some text heuristics that can improve the prediction accuracy indicators?

To answer question 1, we ran a Multi-Layer Artificial Neural Network, trained by the Twitter hate-speech dataset, and tested by our insulting-non insulting dataset. The Twitter hate-speech dataset used for training, included 24k tweets labeled as hate speech, offensive language, or neither, as a training set, and we filtered 9526 sentences which were labeled as insulting/offensive or non-insulting, without difference of opinions between the rankers. As a result, the multi-layer neural network, successfully used later in our experiments, was able to predict the insulting sentences only with a precision of 60%, a recall of 21% and an F1-score of 31%. This result shows the challenge of learning insulting conversation sentences, and in particular, the fact that learned terms from on-line resources of users text cannot be the only source of detecting insulting talk in real world speech and conversation.

To answer question 2, we divided the insulting/non-insulting dataset into a training set and a test set, where 90% of the collected sentences randomly chosen were used as a training set and the rest of the sentences were used as a test set. Our results are provided and explained below. In particular, most of the machine learning methods we used, successfully predicted the label (insulting/non-insulting) of the sentences in the test-set, with a recall, F1-score and accuracy of more than 75%, as presented in Table I. However, if we append the training set with the Tweeter hate-speech sentences described above, the precision is not improved, but the recall of the results decreased to 63%. Thus adding the additional data from Tweeter lowered the percentage of the detected insulting sentences from over 75% to only 63% in total. Consequently, we proceed by concentrating solely on our insulting/non-insulting dataset.

To answer question 3, we compared different models trained on 90% of our dataset and tested on the rest. We examined the following machine learning models: a Multi-Layer Artificial Neural Network (ANN) with five hidden layers, each including 100 neurons; A Tree-Bagger based on a voting procedure with 100 decision trees, and SVM. The maximum number of iterations was determined as 5000 for the Multi-Layer ANN and for the SVM. The methods implementation were imported from Scikit-Learn [2]. In all methods we used, the training sentences were first vectorized, and transformed to TF-IDF vectors, and then sent to the different machine learning methods. Table I describes our results, where 100 random divisions of the data to training and testing set were used. The precision, recall and F1-score results are the mean values of the 100 obtained results.

TABLE I
COMPARISON OF THE CLASSIFICATION METHODS FOR PREDICTING INSULTING SENTENCES

Method	Precision	Recall	F1 Score	Accuracy
Multilayer ANN	0.766	0.772	0.768	0.768
Naive Bayes	0.744	0.726	0.734	0.738
SVM	0.749	0.809	0.777	0.769
Decision Tree	0.737	0.731	0.733	0.735
Tree Bagger	0.755	0.776	0.765	0.762

As shown in these results, the Multi-Layer ANN and the Tree-Bagger method reached precision, recall and F1-score between 75%-80%. This gives us a promising method that can assist the automated agent in recognizing insulting sentences in real-world conversations. The SVM method reached an even higher recall score (80.8%), but its precision value was slightly lower than that of the Multi-Layer ANN and Tree-Bagger. (74.9%).

Finally, in order to answer question 4, we checked whether the former results specified in Table I can be improved by some textual preprocessing. The motivation for this step lies in the fact that only words which appear ten or more times in the dataset were considered parameters in the learning process, so words with a clear sentiment meaning (positive or negative) may exist and appear in our dataset less times than the threshold , but should still be considered in the learning process. Consequently, we imported a list of positive and a list of negative words, introduced by [8][3].

In order to consider the positive and negative words, we ran the following process. For each sentence that includes a positive recognized word, we added the word ”positive” to the sentence, and for each sentence that includes a word from the negative list, we added the word ”negative” to the sentence. Nonetheless, we did not add the word ”like” at all, since, despite the fact that it is listed as positive, it appears several times also in an insulting sentences, such as ”You look like....” or ”You behave like....”. This preprocessing slightly improved our results in most of the cases, as depicted in Table II.

The above results are impressive, given the fact that insulting content detection was based on a single sentence,

[2]http://scikit-learn.org/

[3] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

TABLE II
COMPARISON OF CLASSIFICATION METHODS, WITH EXPLICIT DETECTION OF POSITIVE/NEGATIVE WORDS

Method	Precision	Recall	F1 Score	Accuracy
Multilayer ANN	0.768	0.772	0.769	0.77
Naive Bayes	0.742	0.726	0.733	0.737
SVM	0.77	0.804	0.786	0.783
Decision Tree	0.756	0.749	0.752	0.755
Tree Bagger	0.778	0.779	0.778	0.779

without any additional available information. However, a real world automated assistant agent, that is built to assist children with ASD, can use the textual method described in this study, combined with information from speech signals [6], identity of the speaker or listener, etc., in order to obtain a more accurate prediction, and to be able to notify the child when detecting potentially insulting content.

IV. CONCLUSION

In this study, we composed a special dataset, based on reports of parents of children with ASD, which consists of insulting and non-insulting sentences. We tested the abilities of different machine learning techniques to predict the insulting sentences of the test set sentences based on the trained sentences in the training set. We found that the the best predictors among the machine learning methods were the SVM method, with 80% recall and precision of more than 75%, and the Multi-Layer Neural Network and the Tree Bagger, with precision and recall of more than 75%. Our results can be used to develop an automated agent that will be aware of the special child's social interactions, will detect insulting sentences he says unintentionally or insulting sentences said to him, and will be able to suggest appropriate responses.

In future work, we intend to add additional knowledge in order to be able to recognize sentences that may be insulting or not insulting, depending on their context. In addition, we would like to develop automated tools to assist the child with appropriate responses (or apologies) after recognizing the insulting sentence. Moreover, we would like to focus on emotion recognition from speech. For this task, additional parameters can be considered, such as the tone and pitch, the context of speech, etc. The ability to recognize emotions can increase the ability of the assisting agent to detect the insulting sentences by recognizing the feelings of the insulted listeners according to their responses, in addition to the sentence classification performed according to the text. The combination of the capabilities of detecting insulting sentences via their text and via the voice of the listeners' responses can be exploited to develop a reliable automated agent which will assist the child improve his social relations and functioning.

V. BIBLIOGRAPHY

REFERENCES

[1] A. Abbasi, H. Chen and A. Salem, Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums, ACM Transactions on Information Systems, 26 (3), (2008).

[2] American Psychiatric Association . Diagnostic and statistical manual of mental disorders (DSM-V). Arlington, VA: American Psychiatric Publishing.(2013)

[3] Baron-Cohen, S., Leslie, A. M. and Frith, U. Does the autistic child have a theory of mind? Cognition, 21, 3746.(1985)

[4] S. Berggren, Emotion recognition and expression in autism spectrum disorder:Significance, complexity, and effect of training, PhD diss., the Department of Womens and Childrens Health Karolinska Institutet, Stockholm, Sweden, (2017).

[5] Sofiane Boucenna, Antonio Narzisi, Elodie Tilmont, Filippo Muratori, Giovanni Pioggia, David Cohen and Mohamed Chetouani, Interactive technologies for autistic children: A review, Cognitive Computation, Volume 6, Issue 4, pp 722740, (2014).

[6] Lijiang Chen, Xia Maoa, YuliXue and Lee Lung Cheng ,Speech emotion recognition: Features and classification models, Digital Signal Processing, Volume 22, Issue 6, Pages 1154-1160, December 2012.

[7] Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu and Yu Zhou, Event-Driven Emotion Cause Extraction with Corpus Construction, The Conference on Empirical Methods on Natural Language Processing 1639-1649. (2016).

[8] Minqing Hu and Bing Liu, Mining Opinion Features in Customer Reviews, Proceedings of Nineteeth National Conference on Artificial Intelligence (AAAI-2004), San Jose, USA, July (2004).

[9] Hughes C. and Leekam S. What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. Soc Dev. 13(4): 590 619.(2004)

[10] Kai Gao, Hua Xu and Jiushou Wang, A Rule Based Approach to Emotion Cause Detection for Chinese Micro-Blogs, Expert Systems with Applications (4517-4528), (2015).

[11] Vishal A. Kharde, S.S. Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications, 139 (11), (2016).

[12] Kumar Ravi a,b, Vadlamani Ravi A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications, Knowledge-Based Systems 89, 1446. (2015).

[13] Leslie AM, Frith U. Autistic childrens understanding of seeing, knowing, and believing. Brit J Dev Psychol.; 6: 31524. (1988)

[14] Maano, C., Normand, C. L., Salvas, M.-C., Moullec, G. and Aim, A. Prevalence of School Bullying Among Youth with Autism Spectrum Disorders: A Systematic Review and Meta-Analysis. Autism Research, 9(6), 601615.(2016)

[15] Chikashi Nobata, Joel Tetreault, Achint Thomas Embibe, YasharMehdad and Yi Chang , Abusive Language Detection in Online User Content, WWW '16 Proceedings of the 25th International Conference on World Wide Web, Pages 145-153. (2016).

[16] Tager-Flusberg, H. Evaluating the theory-of-mind hypothesis of autism. Current Directions in Psychological Science, 16, 311315. (2007)