Full Length Article

# Automatic analysis and identification of verbal aggression and abusive behaviors for online social games

Koray Balci *, Albert Ali Salah

Computer Engineering Department, Bogazici University, Istanbul, Turkey

A B S T R A C T

Online multiplayer games create new social platforms, with their own etiquette, social rules of conduct and ways of expression. What counts as aggressive and abusing behavior may change depending on the platform, but most online gaming companies need to deal with aggressive and abusive players explicitly. This usually is tied to a reporting mechanism where the offended player reports an offense. In this paper, we develop tools for validating whether a verbal aggression offense report refers to a real offense or not, in the context of a very popular online social game, called Okey. Our approach relies on the analysis of player behavior and characteristics of offending players. In the proposed system, chat records and other social activities in the game are taken into account, as well as player history. This methodology is sufficiently generic, and it can be applied to similar gaming platforms, thus describing a useful tool for game companies. We report our results on data collected over a six months period, involving 100,000 users and 800,000 game records, and illustrate the viability of such analysis, while providing insights on the factors associated with verbal aggression and abusive behavior for social games.

© 2014 Elsevier Ltd. All rights reserved.

## 1. Introduction

Online social games provide rich interaction possibilities to their users, and create micro-worlds with social rules that parallel, but do not completely overlap with the real world. Since most transactions and interactions happen over digital media, these platforms present great opportunities to analyze user behavior. In online social games it is possible to record user actions, to create or to filter target interactions, and to obtain contextualized behavior instances. With the help of these data, one can either improve the game experience, by for instance adapting the game to maximize player enjoyment (Asteriadis, Shaker, Karpouzis, & Yannakakis, 2012), or use the game for a better understanding of the players themselves, for instance by inferring personality traits from in-game behavior (van Lankveld, Spronck, van den Herik, & Arntz, 2011).

There is a significant body of work that investigates the effects of aggressive and violent content in computer games on the players, particularly whether violent games induce aggression in children or not (Egenfeldt-Nielsen, Smith, & Tosca, 2013; Griffiths, 1999). However, little research has been done on aggressive behaviors within computer games. We do not deal here with the controversial issues of violent games (Ferguson, 2013). We distinguish here

avatar aggression, which involves aggression displayed by the virtual characters of a game, from **player aggression**, which implicates the actual player as the target of aggression. The latter is a form of cyber-aggression, and is often disruptive for gaming experience. In this paper, we deal specifically with verbal player aggression via in-game communication channels. Most social online games provide several communication channels, including in-game chat, private messaging, gifting (i.e. sending a virtual gift to another player), message boards, friendship and alliance requests, and such. Rapid identification and resolution of verbal aggression over these channels is important for the gaming community. For this purpose, the content of verbal messages should be analyzed automatically.

In addition to verbal messages, we explore in this work a number of features that can be used for player profiling in social online games. In particular, we use a supervised machine learning approach to create models of abusive and aggressive verbal behavior from labeled instances of abuse in such an online game, based on actual player complaints. While mechanisms for handling player complaints exist in most social games, game moderators need to spend time and energy to analyze player complaints to resolve each case individually.[1] Subsequently, labeled data are

* Corresponding author at: Aytar Cad. No: 20 D: 12 Levent, Besiktas, 34340 Istanbul, Turkey. Fax: +90 212 287 11 91.
   E-mail address: balci@cmpe.boun.edu.tr (K. Balci).

---

[1] One of the bigger *Okey* sites, run by MyNet (https://apps.facebook.com/canakokey/) has over 1 million monthly active users, and reported receiving about 40 player complaints per hour on the average. Four full-time staff members are hired to deal with these complaints.

costly to obtain. We introduce here a labeled corpus for this purpose. Our study aims to improve the game experience indirectly, by automatically analyzing player complaints, and thus helping game moderators to respond to aggressive and abusive behaviors in the game. At the same time, our analysis may contribute to a better understanding of the factors that underlie such behaviors.

Our approach is based on the analysis of player complaints, player behavior, and player characteristics, including demographic data, game play statistics, and similar features of player history. The social interactions we analyze include chatting, as well as in-game friendship, offline messaging, and gifting. Our profiling methodology performs with a small number of false positives, and is now being incorporated into an actual game environment.

Our evaluation is based on the performance analysis of the classifiers we build for detecting abusive verbal behaviors automatically; if a classifier can perform well, this means the features we look at are selected correctly.

For classification, we have used the Bayes Point Machine formalism in this work (Herbrich, Graepel, & Campbell, 2001). To evaluate our proposed methodology, we have collected the CCSoft Okey Player Abuse (COPA) Database over six months of game play, with 100,000 unique users, and 800,000 individual games. Our labeled complaint data comprises 1066 player complaints, each involving one or more game plays between involved players.

The main research questions at the onset of this study were about understanding how social performance and gaming behaviors relate to verbal aggression, and whether there were factors that correlate highly with verbal aggression and abuse, or common features of abusive players. Our hypothesis is that player profiling and analysis of gaming behavior can provide useful cues in assessing cases of verbal aggression. In addition to answering these questions in the context of a particular game, we have sought to create an application of practical value, to help game designers in the moderation of their online game communities.

This paper is organized as follows. First, we give an overview of related work in social game analytics and aggressive behavior detection. We then introduce the game of *Okey* used in our experiments, and describe its social role in the Turkish culture. Next, we explain our proposed methodology. We present the COPA database, describe its annotation, and report experimental results. We conclude with a summary of findings and limitations.

## 2. Related work

A recent survey on human behavior analysis for computer games illustrates that while game designers analyze player behavior intensively when designing their games, real time behavior analysis is rarely incorporated into the game (Schouten, Tieben, van de Ven, & Schouten, 2011). There are companies that adapt their game content to user preferences by means of A–B testing, where a group of users receive one version of the game, while a second group receives a slightly modified version, and the preferences are recorded to select one of the versions over the other. Gaming companies that govern online games with many subscribers also use data analysis tools in monitoring player activity, for instance to detect cheating behaviors (Yan & Choi, 2002), or for the analysis of player performance in different dimensions like demographics, archetypes, classes, and sub-classes (Shim, Sharan, & Srivastava, 2010). These tools, also called game analytics, have direct impact on game revenues, and therefore are receiving more and more interest (El-Nasr, Drachen, & Canossa, 2013; Mellon, 2009). The system we propose in this paper can be seen as such an analysis tool to help the governance of an online social game.

Game analytics have been used previously for detecting different types of players. For instance Drachen, Canossa, and Yannakakis (2009) have used unsupervised learning techniques

on game analytics data to cluster the players into four groups, according to gameplay. There have been several studies in the literature that demonstrated the usefulness of gaming platforms for inferring behavioral cues about the players. Yee, Ducheneaut, Nelson, and Likarish (2011) conducted a survey study on 1040 *World of Warcraft* (*WoW*) players, focusing on demographics and personality, and correlated the survey results with four months of game playing data. The authors were able to verify some expected results; for instance Extraverts as determined by the personality survey indeed preferred group activities over solo activities. This indicates that in-game behavior may be correlated with actual player behavior. More recently, Bean and Groth-Marnat (2014) conducted a similar study with 1210 *WoW* players to examine the connection between personality characteristics of the players measured by the 44-item personality measure *Big Five Inventory* (*BFI*) (John, Donahue, & Kentle, 1991), and the playing style and personal profiles of the players. They report correlations between these, but found no support for antisocial behavior or aggressiveness in relation to the personality scores of the players when compared with the markers of antisocial personality factors (Markey & Markey, 2010).

In another study, van Lankveld et al. (2011) designed a game module for *Neverwinter Nights*, and through carefully tailored interaction options, used it to measure user behavior. Using correlation analysis on 275 game variables, they obtained relationships with five personality traits and the video game data. In both studies, the Big Five personality traits survey was used (Digman, 1990).

The gaming behavior we study in this paper involves multiparty chat messaging (among other variables). Multiparty chat refers to communications in microtext format where multiple participants converse asynchronously via text messages. Uthus and Aha (2013) provide a survey of artificial intelligence methods applied to the analysis of multiparty chats, and establish that while multiparty chat analysis has been the focus of substantial research in social and behavioral sciences, very few studies have been conducted for chat analysis in the gaming context. While we do not analyze the actual chat content to a great depth in this study, our study contributes to the field through the inclusion of non-verbal signals and bad language usage in our analysis.

Our focus in this work is on verbal aggression in social games. Verbal aggression is one of the major categories of violent behavior (Yudofsky, Silver, Jackson, Endicott, & Williams, 1986). In the literature, it is measured with the help of different scales and inventories (Archer, Kilpatrick, & Bramwell, 2006). While most aggression research focuses on factors related to physical and verbal aggression, little is known about online aggression, except that Internet harassment related literature (including unwanted sexual solicitation) is surprisingly large. However, there exists controversial points of views on the surge of Internet harassment (Finkelhor, 2014).

Abuse and aggression in online social games come in different forms, and can be studied under the umbrella term of *cyber aggression* (France, Danesh, & Jirard, 2013). A well-studied form of cyber aggression is cyberbullying (Li, 2007). Cyberbullying usually refers to prolonged mistreatment (Kwan & Skoric, 2013), whereas in the application we discuss here, abusive behaviors can also happen once. Reynolds, Kontostathis, and Edwards (2011) have proposed to use text-mining techniques for automatically detecting cyberbullying from Internet posts. This work resembles our approach, but relies exclusively on textual content, whereas we put the stress on historical factors to determine prior probabilities of exhibiting abusive behavior.

In the next section, we describe a social online game called *Okey* that relies heavily on multiparty chats for social interaction. To our knowledge, there has been no previous studies on this well-known Turkish game, excepting the preliminary study that precedes this paper (Balci & Salah, 2013).