PLAYER PROFILING AND ANALYSIS OF ABUSIVE BEHAVIOUR IN SOCIAL GAMES

by

Mehmet Koray Balcı

BSc, in Electrical and Electronics Engineering, Middle East Technical University, 1998

MSc, in Cognitive Sciences, Middle East Technical University, 2002

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in
Boğaziçi University
2015

ACKNOWLEDGEMENTS

This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 114E481.

This work is supported by the Turkish Ministry of Development under the TAM Project number DPT2007K120610.

I thank Rıdvan Salih Kuzu and Eda Aydın for their contributions.

Abstract

PLAYER PROFILING AND ANALYSIS OF ABUSIVE BEHAVIOUR IN SOCIAL GAMES

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. Artificial intelligence and machine learning techniques are not only useful for creating plausible behaviors for interactive game elements, but also for the analysis of the players to provide a better gaming environment.

In this thesis, we investigate the verbal and non-verbal data generated in an online social gaming platform and propose novel algorithms for automatic classification of abusive players and player complaints. We use features that describe both parties of the complaint (namely, the accuser and the suspect), as well as interaction features of the game itself. This methodology is sufficiently generic, and it can be applied to similar gaming platforms, thus describing a useful tool for game companies.

We also introduce the COPA Database of 100.000 unique users and 800.000 individual games, which includes multiparty chat records in Turkish, in addition to player profiles, social interactions, and annotated complaint data. The proposed supervised methodologies for complaint classification are tested on this database, and we advance the state-of-the-art in this challenging problem. In addition, we have studied the multiparty chat data collected within the COPA dataset. In particular, we developed a methodology for affect analysis to enrich the interpretation of the data. Finally, we developed a system for authorship recognition based on chat records to identify duplicate user accounts and returning abusive users by analyzing the chat data.

ÖZET

SOSYAL OYUNLARDA OYUNCU PROFİLLEME VE TACİZ DAVRANIŞLARININ ANALİZİ

Çevirimiçi çoklu oyunculu bilgisayar oyunları, kendine özgü davranış kuralları ve sosyal normları ile yeni sosyal platformlar oluşturmaktadır. Bulunulan platforma göre saldırgan ve tacizkar davranışlar değişkenlik gösterebilmesine rağmen, hemen hemen tüm oyun üreticileri bu tür davranış ve istenmeyen durumlar ile aktif olarak mücadele etmektedir. Yapay zeka ve otomatik öğrenme yöntemleri, geliştiriciler tarafından oyun öğelerinin etkileşiminde kullanılmasının yanında daha sağlıklı bir oyun ortamı oluşturabilmek için oyuncuların davranışlarının analizinde de kullanışlı olabilmektedir.

Bu tez çalışmasında, çevirimiçi sosyal bir oyun platformunda toplanmış oyuncuların sözel ve sözsüz iletişim verisi üzerinde incelemeler yaparak tacizkar oyuncular ve ilgili oyuncu şikayetlerinin otomatik olarak sınıflandırılması üzerine yenilikçi yaklaşımlar öneriyoruz. Çalışmamızda oyuncu şikayetlerinde taraf olan oyuncuların (şikayet eden ve edilen) oyun profillerinden oluşan öznitelikler ile birbirleri arasında geçen iletişime dair bilgiler üzerinden çıkarımı yapılan öznitelikleri kullanmaktayız. Bu yenilikçi yaklaşımın benzer özelliklere sahip başka oyun platformlarına adapte edilerek oyun geliştiricilere yardımcı olacağını öngörüyoruz.

Çalışmamızda kullandığımız 100.000 tekil kullanıcı ve 800.000 oyun verisi barındıran COPA oyun veritabanını sunuyoruz. Bu veritabanı, Türkçe çoklu sohbet verisinin yanında oyuncu profili, sosyal etkileşimleri ve elle incelenmiş ve işaretlenmiş oyuncu şikayetlerini içermektedir. Önerdiğimiz yöntemler bu veritabanı üzerinde deneylerden geçirilmiştir. Son olarak, çoklu sohbet verisi kullanarak geliştirdiğimiz iki ayrı çalışmayı sunmaktayız. Bu çalışmalarda, öznitelik kümemizi iyileştirme amacıyla otomatik duygu analizi ve aynı oyuncuya ait birden fazla oyuncu hesabı olup olmadığını kestirebilmek

için geliştirdiğimiz yöntemleri anlatmaktayız.

Contents

List of Figures

List of Tables

LIST OF SYMBOLS/ABBREVIATIONS

1. Introduction

1.1. Problem Definition

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 analyse 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 [1], or use the game for a better understanding of the players themselves, for instance by inferring personality traits from in-game behavior [2].

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 [3, 4]. However, little research has been done on aggressive behaviors within computer games. We do not deal here with the controversial issues of violent games [5]. 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 study, 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.

1.2. Research Questions

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.

1.3. Approach of the Thesis

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. 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 manual mechanisms for handling player complaints exist in most social games, game moderators need to spend time and energy to analyse player complaints to resolve each case individually[1] .

Subsequently, labeled data are costly to obtain. We introduce here a labeled corpus for this purpose. Our study aims to improve the game experience indirectly, by automatically analysing 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.

The gaming behavior we study involves multiparty chat messaging among other variables. Multiparty chat refers to communications in microtext format where multiple

[1]One of the bigger Okey sites, run by MyNet (https://apps.facebook.com/canakokey/) in Turkey 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.

participants converse asynchronously via text messages. Uthus and Aha 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 [6]. Reynolds et al. previously used machine learning to detect language patterns that indicated cyberbullying [7]. 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.

For analysis of chat content, we utilize non-verbal, language independent features and show that these contribute to overall performance of abusive player and complaint classification task. Then, we present a study on verbal communication and affect analysis that use the same chat dataset. Moreover, we show that an offensive player can be recognized even if that player creates a new account by only analyzing and comparing the chat input.

Our study has further motivational roots in the literature of personality computing in psychology [8]. Close scrutiny of our data shows that some of the conflicts between players arise from the misunderstanding of expressed social cues. The *Brunswik Lens* model has recently gained importance in multimedia computing, and provides a useful abstraction for social interactions [9]. This concept helps one to conceptualize complaints as composite constructs, including the social cues emitted by one party, and the percept created by the other party based on inference on these cues [10]. Subsequently, the Brunswik Lens suggests that the analysis of a complaint (as well as other social constructs) should include both the source and the receiver of the social message, in our case, the *accuser* and the *suspect*, respectively.

Our focus in this work is mainly on verbal aggression in social games. Verbal aggression is one of the major categories of violent behavior [11]. In the literature, it is measured with the help of different scales and inventories [12]. 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 [13].

Abuse and aggression in online social games come in different forms, and can be studied under the umbrella term of *cyber aggression* [14]. A well-studied form of cyber aggression is cyberbullying [15]. Cyberbullying usually refers to prolonged mistreatment [16], whereas in the application we discuss here, abusive behaviors can also happen once. [7] 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.

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 analyse 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.

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 (e.g. due to use of hate speech, insults, aggressive behavior, etc.) exist in social games, game moderators need to spend time and energy to analyze player complaints to resolve each case individually. Subsequently, labeled data are costly to obtain. We have previously introduced the labeled COPA corpus for this purpose [17], and proposed an approach to classify abusive players by evaluating player profiles and in-game data [18]. More recently, in our follow-up study [19], instead of classifying players, we directly investigate and classify individual complaints, which may involve abusive or offensive behavior. We also take into account the ac-

cusing player's in-game data, as well as both accuser and suspected players' recent communication history with each other. This not only doubles the number of features we use compared to our previous study (19 features in [18] as opposed to 43 features in the present work), but also fundamentally changes the perspective taken. It is not only the offender's data that should be investigated to judge cases of abuse, but also the accuser, as well as the interaction between them. We report in our experimental evaluations the clear improvement of the proposed approach in comparison with the approach that just looks at the offender.

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.

This indirect evaluation is also a result of the assumption that abusive players could not be expected to cooperate in a direct evaluation, for instance by filling surveys about their behavior.

1.4. Structure of the Thesis

Chapter 2.1 describes the main approaches used by major game companies to regulate player behavior in their online platforms. This is typically achieved by maintaining moderation teams, and by providing guidelines and rules for the players.

Chapter 2.2 describes what analytics and game analytics are and overviews game analytics approaches employed by gaming companies. We also compare analytics with manual user surveys in this chapter.

Next, we present our survey on player aggression and abusive actions in Chapter 2.3. We also highlight some of the data driven approaches on game related studies in our overview of the literature.

To evaluate our proposed methodology described in detail in Chapter 4, 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 1.066 player complaints, each involving one or more game plays between involved players. COPA database and Okey Game are introduced in Chapter 3.

In addition to our main focus on player and complaint classification, we present our studies on verbal communication analysis in Section 4.5 and authorship recognition in Section 4.6. Then, we present an overview of the machine learning methods we studied and experimented with in Section 4.7.

Our experimental results are laid out in detail in Chapter 5. Finally, we conclude this study in Chapter 6 with discussion of results and limitations on abusive player and complaint classification.

1.5. Contributions

During the course of this study, we have published two journal and three conference papers.

In 2013, we presented our preliminary results and introduced the COPA dataset at DESVIG workshop in CHI conference [17].

Our more elaborate study on automatic classification of abusive players was published in Computers in Human Behaviour journal in early 2015 [18]. We describe our approach on this topic in Section 4.3.

Next, we published our enhanced results on complaint classification with an enriched feature set in IEEE Transactions on Computational Intelligence and AI on Games in late 2015 [19]. Methodology of this study is laid out in Section 4.4.

For the main focus of this thesis mentioned above, we present our detailed experimental results in Section 5.

Following the above work authored by the author and supervisor of this thesis, we worked with Eda Aydın Oktay in the scope of her MSc studies at Computational Science and Engineering to analyze verbal communication further and published a conference article in Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies in 2015 [20]. This study is explained in detail in Section 4.5.

More recently, our work on authorship recognition presented in Section 4.6 is accepted for publication in proceedings of International Workshop on Biometrics and Forensics (IWBF) 2016. This is also a joint work with MSc candidate Rıdvan Salih Kuzu in System and Control Engineering.

2. Related Work and Background

2.1. Player Behavior Regulation in Multiplayer Online Gaming

Our proposed framework to evaluate player complaints and classify abusive players target games that are played online and offer social interaction among the players. Note that the term "social games" is commonly used for a broad range of game genres[2] that allow or require social interaction between players [21]. There are several game genres that offer multiplayer activity in which players can connect and share the same game environment together and collaborate or compete against each other in the game world. Before proceeding, we will overview some online game genres that can benefit from our study.

Online counterparts of card games and board games are a good example for social games. Poker, Backgammon, Okey, Monopoly, Scrabble and other well known card and board games in real life have several popular online versions. These games usually attract people of all ages, and can be classified under a more broader group of games called "casual games". Casual games require no long-term commitment, where people can join for a reasonably short amount of time and enjoy their time. Almost all versions of these games include some form of player interaction in order to enhance player engagement and loyalty and attract a larger audience.

Massively multiplayer online games (MMOs) on the other hand, require a fair amount of dedication and experience in order to succeed in. Different game genres such as strategy games, role playing games and action games have MMO versions. In addition to dedication, these games usually require skills such as dexterity, reflex, focus, fast problem solving, strategic thinking as well as collaboration and teamwork. Some of these games have also promoted professional competition where players or teams of players compete in events that offer large prizes. Professional gamers are sponsored

[2]For an exhaustive list of game genres, see `https://en.wikipedia.org/wiki/List_of_video_game_genres`

by international companies and national teams are formed in worldwide events that are followed by big audiences. To illustrate, League of Legends World Championship 2015[3] offered a prize pool of 2,130,000 USD, took place in Paris, London, Brussels and Berlin, in some of the largest stadiums in Europe as shown in Figure 2.1. In recent years, e-gaming and electronic sports have matured with an international federation that regulates events, rules, professional players and teams as well as sponsors just like any other industrial sports association.

Figure 2.1. League of Legends World Championship 2015, image taken from `http://www.lolesports.com/en_US/worlds/articles/everything-you-need-know-about-worlds-2015-knockout-stage`

Although MMOs usually require extra dedication and enhanced skills, they also contain means to communicate and socialize within the game environment. In most MMOs, players can form a team and play against each other. The story lines of these games have a vast variety, from ancient historical settings to first and second world wars to science fiction.

In this section, we will give an overview of major online multiplayer games and

[3] `http://lol.gamepedia.com/2015_World_Championship`

how they respond to abusive and unwanted behavior. In each subsequent section we will briefly present each game and explain how they treat such incidents. In summary, every game we studied, there are some means to warn users on prohibited behaviours and actions, mechanism to report such actions and certain types of punishments that may be given to offending players. The only exception to this standard approach is the *League of Legends (LoL)* game from Riot Game Studios, such that they have innovative means of dealing with unwanted actions. We will focus and discuss the methods employed by LoL team in more detail.

Most games we will overview have a revenue model commonly known as free-to-play. One can create an online account and play these type of games without paying real money. However, these games generally contain sold items in order to create revenue such as paid subscriptions with enhanced experience, sold virtual game items or extra features only accessible when players pay. Free-to-play model usually attracts large crowds and once a game becomes popular, even a small percentage of the subscribed players can produce substantial amounts of income for these companies.

Blizzard Entertainment[4] is one of the most successful game development companies of all times with major titles such as *World of Warcraft (WoW)*, *Diablo* and *Starcraft* franchises. As one of the first massively multiplayer online (MMO) game developers with their Battle.net online platform dating back to 1997, Blizzard Entertainment is treated as a pioneering studio in games industry with high quality and highly lucrative products. In Battle.net platform, players can create an account and join several Blizzard games. The servers of these games are maintained by Blizzard Entertainment and to play most of the games on Battle.net platform, players have to pay monthly fees.

WoW game itself is a massively multiplayer online role playing game (MMORPG) where players can choose a character to play and want to impersonate. They can create a character by selecting gender, race, visual appearance, individual skill levels for strength, dexterity, stamina, agility, wisdom, etc. Each character in a role playing

[4] `http://blizzard.com`

game (RPG) that has a story line of fantasy world as WoW belong to a character class such as warrior, magician, healer, etc. These classes narrow down the type of skills each character can have and advance and creates diversity in the game world. WoW players form teams to participate in various game quests with small narratives. Each quest involve several battles with non-player (AI driven) characters and other teams. When a team wins a battle, they gain experience which contribute to how fast they level up and advance within the game and become more powerful. Winners also are awarded by valuable items which will enhance their future battles. Usually, after battles, players in a team discuss what went well and wrong. They can also communicate with each other while the game is in progress, to form a strategy and motivate each other. However, in some cases, players can harass or abuse one another. These cases are usually reported to the game moderators by the victimized players.

Figure 2.2. A screenshot from World of Warcraft

Blizzard Entertainment has a set of general rules for in-game and forum conduct on Battle.net platform.[5] Violations of the conduct may result in warnings, suspensions, or permanent account closure by the moderators of the platform. The conduct defines minor, serious and severe violations where *Harassment or Defamation* is listed as a minor violation and explained as *insulting other characters, players, Blizzard Entertainment employees, or groups of people in-game and externally.*

[5] `https://us.battle.net/support/en/article/ingame-and-forum-conduct`, accessed: 18-10-2015

Obscene or Vulgar Language is a type of serious violation and defined as *"using language that is crude, offensive, or pornographic in nature and/or referring inappropriately to human anatomy or bodily functions"*.

Severe violation category includes several actions relevant to our study:

- Threatening someone with real-life violence or harm.
- Distributing another player's, or a Blizzard employee's real-life information.
- Promoting racial, ethnic, or national hatred.
- Using a racial, ethnic, or national slur.
- Referring to extreme and/or violent sexual acts.
- Referring to extreme and/or violent real life actions.
- Insulting or negatively portraying someone based on their sexual orientation.
- Negatively portraying major religions or religious figures.
- Insulting or negatively portraying someone based on their religious orientation.
- Alluding to symbols of racial, ethnic, or national hatred.

They note that the above violations, especially real-life threats or distribution of personal information can have serious consequences such that the company can engage local law enforcement to ensure the safety of all parties involved.

Depending on the severity of the violation, game moderators can apply the following penalties:

- Text warning
- Temporary account suspension
- Account closure

In their guild chat environment, which is a voluntary in-game platform that players can choose to participate or not, Blizzard Entertainment punishes only the most severe violations (i.e. racial/ethnic/national, extreme sexual/violence, real-life threats) in case of complaints from other players.

In 2010, the company announced *Real ID*, which would require users to login and post to forums with their real name to prevent unwanted behaviour. However, response from their user community was totally negative, taking it as an attack to their personal privacy. Although some users argued using real names would enhance the tone in the forums, most worried such an action may have dangerous real life consequences such as stalking, harassment, and employment issues.[6] Due to strong negative response, Blizzard Entertainment later took a step back and announced they would not enforce *Real ID*.

Call of Duty (CoD) is a first person shooter online video game franchise, which was first released in 2003. The franchise consists of more than 20 games released so far and as of April 2015, the Call of Duty series has sold over 175 million copies.[7] The game simulates the infantry and combined arms warfare of World War II. Multiplayer mode usually consists of up to 32 players split into two teams, combating to win points for various game modes such as "Capture the Flag", "Deathmatch", "Search and Destroy", etc. Team players have the means to chat before, during and after each game session. CoD games are also used in e-gaming competitions and major league gaming, with trophies including cash prizes to winners.

In terms of community and player management in CoD series, here we give an example from one of the games in the series, namely "Call of Duty: Black Ops 2". The security and enforcement policy of the game[8] states that all infractions undergo a thorough review process by the company security team before enforcement, and penalty is not subject to further review. Policy section for offensive behaviour is stated as follows:

> Any user who is found to use aggressive, offensive, derogatory or racially charged language is subject to penalty.
> - First offense: User will be temporarily banned from playing the game online.

[6] `http://www.gamasutra.com/view/news/120220/InDepth_Why_Was_Blizzards_Real_ID_Such_An_Issue`

[7] `http://www.gamezone.com/news/call-of-duty-franchise-surpasses-175-million-copies-sold-3414623`, accessed: 18-10-2015

[8] `https://support.activision.com/articles/en_US/FAQ/Call-of-Duty-Black-Ops-II-Security-Enforcement-Policy`, accessed: 18-10-2015

Figure 2.3. A screenshot from Call of Duty: Ghosts

- Second offense: User will be temporarily banned from playing the game online and will have voice chat privileges in the game revoked.
- Extreme or repeat offenses: User will be permanently banned from playing the game online, will have their stats & emblems reset, and will be blocked permanently from appearing in leaderboards.

TERA Online is a massively multiplayer online role playing game (MMORPG). The game has a theme of 3D fantasy storyline and includes questing, crafting and player vs player online action. The game is first released in South Korea and attracts gamers all around the world since with its free to play model available since 2013. The game reached 1.4 million subscribers worldwide.[9]

In their online code of conduct page, offences that are not allowed in the game are listed.[10] . Here are some of the offences relevant to our studies, taken directly from their list;

- Do not attack someone personally, including their race, gender, religion, nationality, ethics, sexual preference, or personal beliefs. Keep it clean.
- Don't harass, stalk, or purposely do things to make someone else feel uncomfort-

[9] `http://www.gamesindustry.biz/articles/2013-03-20-tera-crosses-1-4-million-after-f2p-switch`, accessed: 18-10-2015

[10] `http://tera.enmasse.com/legal/rules-of-conduct`, accessed: 18-10-2015

Figure 2.4. A screenshot from TERA Online

able.

- Do not post or use graphic sexual, grotesque, or violent language or images. There are enough websites out there for this. We don't need another one.
- Child predation or pornography = BAN! And possibly the authorities get involved, too.

The offending players receive temporary mute ban (inability to interact with other players with voice), temporary account suspension and account closure punishments.

Wargaming.net is a European game development company, which mainly focuses on turn based and real time strategy (TBS/RTS) war themed MMO games. Since 2009, they have released several second world war themed games with free to play model and achieved great success with more than 15 game titles. In November 2011, they announced that one of their *World of Tanks* servers in Russia broke the Guinness Record for the highest number of concurrent players on a single server with 250.000 players logged in to the game at the same time.[11] Although the game is free to play, the company sells in-game goods and premium subscription models. Wargaming.net's announced 75 million subscribers to its flagship game "World of Tanks" in

[11] `http://news.mmosite.com/content/q/2011-12-09/world_of_tanks_sets_a_new_guinness_world_record.shtml`, accessed: 18-10-2015

late 2013.[12]

Figure 2.5. A screenshot from World of Tanks

The following is the part of "World of Tank" game rules relevant for our studies;[13]

- Excessive profanity and inappropriate language is not welcome. It is suggested to ensure that the censor filter is switched on. The censor filter is not an excuse to break the existing game and chat rules, and excessive profanity will still be sanctioned.
- Insults, personal attacks, abuse or harassment are not tolerated on any level.
- Derogatory comments based on race, nationality, religion, culture, sex, or sexual preference are prohibited.
- Allusion of racial or national supremacy, as well as discriminative propaganda on any level is prohibited.
- Distribution of user's personal information without their consent is prohibited.
- Slandering users or posting false information about users in all game chats and channels is prohibited.
- Discussion on, or linking to illegal activities, such as illicit drugs, is prohibited. This includes but is not limited to the linking of, or discussion on, websites

[12] `http://www.develop-online.net/news/world-of-tanks-eclipses-75-million-registered-players/0187343`, accessed: 18-10-2015

[13] `http://worldoftanks.com/en/content/wot_game_rules/`, accessed: 18-10-2015

dedicated to vulgar, racist, abusive, illegal, or any other content prohibited by the EULA, or linking to the resources that contain such advertisement or content.

- Death threats and other threats of violence in real life, directed either against individual users, game masters or administration of the project, are prohibited.
- Discussion of social, religious, political, illegal or other controversial topics that may create offense is prohibited. This includes but is not limited to negative portrayal of religious and political figures is prohibited.

Game management evaluates these and similar issues, and act as follows as mentioned in the same document:

- Wargaming.net reserves the right to evaluate each incident on a case by case basis. The action that may be taken may be more lenient or more severe than those listed under each category.
- Wargaming.net may suspend, terminate, modify, or delete accounts at any time for any reason or for no reason, with or without notice to the owner of the account. Accounts terminated by Wargaming.net for any type of abuse, including without limitation a violation of these rules or the *End User License Agreement (EULA)*, will not be reactivated for any reason.

DotA 2 is a multiplayer online battle arena (MOBA) game with a free to play model. Since its release in 2013, the game achieved great success in terms of popularity. With an impressive all time peak of 1,262,000 concurrent players (Feb 2015) and more than 500,000 monthly average players in 2015, DotA 2 is one of the most popular games of all times.

DotA 2 developers have put in place a rather interesting system to prevent abusive and offending players. If a player quits an ongoing game and leaves other players in an unpleasant situation a few times, or gets reported by other players because of their use of foul language in chat, the player is moved to a "low priority queue". The queue acts like a prison, in which all players are offenders and they are required to play amongst each other and without gaining game experience or in-game valuables for a certain

Figure 2.6. A screenshot from Dota 2

period of time or number of games. The system is said to be semi-automatic and triggers punishment by simply counting early disconnections from games and number of reports against the player submitted by other players.

Valve Corporation, the company who developed and published the game, explains their motivation behind the system in a developer blog post as follows;[14]

> "One of the first things we dug into were the factors that contributed to a player quitting. [...] Losing a bunch of DotA 2 games doesn't seem to cause people to quit. But one thing that did stand out in the data was the amount of negative communication between players. Put simply, you are more likely to quit if there is abusive chat going on in your games."

League of Legends (LoL) is a multiplayer online battle arena (MOBA) game developed by Riot Games and released in 2009. Since then, the game attracted millions of players and established the standards for the MOBA genre and e-sports and e-gaming culture. In terms of subscriber counts and their loyalty, LoL is one of the most successfull online games of all times with over 67 million monthly and 27 million daily gamers as of January 2014.[15]

[14] `http://blog.dota2.com/2013/05/communication-reports/`, accessed: 18-10-2015

[15] `http://blogs.wsj.com/digits/2014/01/27/player-tally-for-league-of-legends-surges/`, accessed: 18-10-2015

Figure 2.7. A screenshot from League of Legends

In terms of player satisfaction, Riot Games is one of the innovative companies with their approaches on dealing with player complaints and offensive players in general.

The following actions are considered as punishable offences by the LoL game environment code:

- Explicit use of hate terms, racial slurs, cultural epithets, etc.
- Players who deliberately and viciously insult other players.
- Repeatedly negative, nonconstructive attitudes.
- Players whose teasing crosses the line, and who persist after being asked repeatedly to stop.

In 2011, Riot Games introduced *The Tribunal* system to evaluate user based complaints and incidents in LoL. The aim of *The Tribunal* is to keep the game community as peaceful as it can be and to control abusive and offensive behaviours. For this purpose, the company chose to use the gamers themselves in dealing with player complaints. When a complaint is raised by a gamer, an experienced player (i.e. one that has spent enough time to have achieved at least level 20 status in the game world) can be assigned to deal with the issue. This player will then analyse the issue by looking

at game chat logs, offender's game statistics and other complaint details, like a judge would do in real life. Note that each case is given to multiple judges. Next, each judge gives his/her verdict, either punishing or pardoning the accused player. The actual verdict is decided by majority voting mechanism among the judges. The punishment can vary as permanent / time ban or mere warning, according to the seriousness of the case at hand.

In order to become and remain a judge, a player should also have a high *justice rating* which is based on how often their personal verdict coincides with the overall vote in each incident. If a judge votes against the consensus constantly, they can lose their judging privileges.

On December 2011, Riot Games released the following Tribunal metrics:

- 1.4% of all players have been punished by the Tribunal.
- Over 50% of all punished players never re-offend.
- 94% of players who receive enough reports to face the tribunal are punished by their peers.
- Average player reports for the average one-time offender: 11
- Number of player reports accrued by the average repeat offender: 70
- Offenders lose games: 24% of offenders are on the winning team. 76% of offenders are on the losing team.
- Offenders make bad teammates: 71% of offenders are reported by their own team. 29% of offenders are reported by the enemy team.
- Over 16,000,000 total votes have been submitted.
- Over 80,000,000 influence points has been rewarded to voters.

The system was disabled in 2014 by Riot Games with no news on when it will be enabled again.[16]

Recently, Riot Games announced that they started working on an automatic sys-

[16] `http://leagueoflegends.wikia.com/wiki/The_Tribunal`, accessed: 18-10-2015

tem that detects bad language, homophobia, racism, sexism, death threats and other forms of excessive abuse. The new system will still be triggered by player complaints but the rest of the process will be fully automated where the offenders will be notified immediately and punished if necessary. Riot will utilize machine learning on the massive data related to interaction among the players to learn at a massive scale what type of behaviors are accepted as OK or not OK in the community, and deliver immediate feedback to the players. Riot team reports that initial tests of the system started on some of their game servers in May 2015.[17]

2.1.1. Conclusion

In this section, we made an overview of the major multiplayer online game titles that are prominent in games industry. All of the games we presented attract large amount of players with millions of subscribers and offer players various methods to interact with each other to socialize and collaborate in the game world. In consequence, each game has their own approach for regulating disputes that arise in these communication channels. For games with so much user bases and consecutively so large revenues, it is an utmost importance to preserve player satisfaction and avoid disturbance of the players with unwanted behaviour. Although each game addresses the importance of dealing with abusive act and harassment and taking immediate action against various offences, neither utilize an automated approach to evaluate or assess player complaints or single out abusive players. Therefore we can conclude that the state of the art in games industry on dealing with complaints is to handle each incident separately, employ human moderators to analyze the case thoroughly and apply penalties where necessary. We argue that an automated system to study player complaints and classify abusive players will greatly help games industry in general.

[17] `http://euw.leagueoflegends.com/en/news/game-updates/player-behavior/new-player-reform-system-heads-testing`, accessed: 18-10-2015

2.2. Game Analytics

2.2.1. What is analytics?

Analytics as a general term can be defined as the process of transforming observational data related to a complex system into insight in order to enhance our understanding of the state of that system and make better decisions regarding actions to improve its performance. In software, this can be summarized as collecting information about the usage of a software in order to enhance overall user satisfaction and improve the software itself [22, 23, 24, 25].

In recent years, with the advances on internet accessibility worldwide, online consumption of software products exceeded their uses by offline desktop counterparts. There is an increasing number of software products that are available online, especially via a web browser. These software solutions are commonly called "Software as a Service" (SaaS). SaaS solutions are installed on a server, can be updated regularly and can be accessed and used by huge number of people over internet. This makes it very easy to gather analytics data from users, analyze this data and update the software iteratively. A large user base can also act as a test bed such that alternative ways of implementing user interaction can be tested by a subset of the actual user base before the final decision is made by the developers.

User driven split testing, generally also referred as A/B testing enables developers to produce two different versions of the same product and serve each of them to a different audience and measure how well each alternative performs [26, 27]. As an example, in an e-commerce web site, the placement or color of buttons that direct the users to purchase things can be A/B tested by serving each alternative to a known subset of users for a period of time. During that time, developers gather analytics data and measure how many users click purchase button in each group. The winning color or position of the button is the one that achieves the most clicks.

In summary, online software products have means to access and gather valuable

information from their users and collect metrics that will guide them to choose which features to implement next. It is important to emphasize that the increase of high availability of web based software solutions and mobile devices led to elevated competition among software products trying to gain market share. In addition to the online revolution of software solutions, collecting the information about how these solutions are used by people has also become a standard procedure in the lifetime of a software product.

Although developers can implement their own analytics tools, middleware web and mobile analytics tools (such as Flurry[18] , Google Analytics[19] , GameAnalytics[20] , and Mixpanel[21]) have also matured, and are generally used to gather and analyse user data.

2.2.2. What is game analytics?

Online games also use analytics as a tool to monitor engagement, measure retention and churn by integrating either third party game analytics tools or develop their own in-house solutions [28, 29].

Once an online game achieves large enough user base to gather data, analytics tools help developers figure out what modifications can enhance the performance metric they are seeking to improve. Most of the time developers are interested in preventing players quitting the game as long as they can or increase the number of paying clients and improve gamer satisfaction.

As an example, in an online game, if the analytics data show that a large number of users quit the game session at a certain quest or puzzle, this can be analyzed further. It can signify that the particular game quest can be too hard to complete, boring or not rewarding. It can even mean that the players are too tired of playing the game for

[18] http://www.flurry.com/
[19] http://www.google.com/analytics/
[20] http://www.gameanalytics.com/
[21] http://mixpanel.com/

an extended period of time and they can be at their limits, lacking extra motivation to continue. In such cases, developers have several solutions for each possible reason above. Additional incentives such as extra prizes or virtual gifts can be offered to players just before that certain point. Alternatively, the difficulty level can be tweaked to make it easier such that players can pass beyond that point. Another option can be to add some joyful content such as some exciting sound effect or interesting animation, etc. Each of the alternatives can be a solution, and it is a hard decision to make given the data from player base only points out a problem but not the reasons behind it. There is no possible convenient way to get back to players who already quit the game and get feedback from them about why they left the game. As a result, given enough resources, developers can implement all the solutions and release each of them to a subset of players access and gather some more data for a period of time. With split testing, they can clearly observe which solutions perform best and make their final decision to release which features to release to the greater game audience. Note that, although releasing all features at once without split testing is also an option, this can also result in confusion and dissatisfaction from players. Releasing all features at once can also be split tested.

In summary, game analytics tools focus on game events, and typically do not incorporate any pattern recognition or machine learning modules that can be customized to benefit game behavior moderation.

2.2.3. User Surveys

In addition to automatic collection of data using analytics tools, user surveys is another means to gather valuable information from users. Before the rise of online games and analytics tools, use of surveys was the standard way to convey user based research in academia. However, the procedure to collect data is cumbersome at the very least. Initially, the researchers have a hypothesis to investigate and need to collect data and support their claim. In order to achieve this, a survey should be prepared and cover all the bases that the research is aiming to investigate. Next, the survey should reach enough gamers. For end users and gamers, often there is no motivation to participate in

these surveys. Most of the time, surveys are accompanied by motivational extras such as in-game gifts or extra course credits for students that participate. Recently, most surveys are filled out online, however this doesn't change the fact that the results of the survey should be sorted and analysed by the researchers. Finally, if the actual results support the initial hypothesis by the researchers, there can be a positive outcome from the study. For an exhaustive overview on web surveys, the issues and approaches, see Couper's seminal work on the subject published in 2000 [30].

To sum up, although use of surveys to collect data is a common method in game research, it is a very tedious manual process that result in only a small fraction of data compared to automated tools used in game analytics. Therefore, in our study we only focus on the latter.

2.3. Player Aggression and Abuse

Artificial intelligence (AI) and machine learning methods have their uses on various domains of gaming for a while now [31]. In a recent study, Yannakakis listed the major game-related domains in which AI techniques found various uses, as well as the interactions of these domains and their mutual influences [32]. While automatic analysis of game complaints is a novel area in game research, it can be positioned in this panorama somewhere close to player modeling [33].

In recent years, web and mobile analytics tools have matured, and are frequently used to gather user data. We present an exhaustive overview of major analytics tools available in Section 2.2. Many games incorporate these third party tools to track user behavior, to monitor engagement, to measure retention and churn. These tools focus on game events, and typically do not incorporate any pattern recognition or machine learning modules that can be customized to benefit game behavior moderation. Nonetheless, there has been a significant amount of research on game data analysis [28, 34, 35].

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 [36]. 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 [37], or for the analysis of player performance in different dimensions like demographics, archetypes, classes, and sub-classes [38]. These tools, also called game analytics, have direct impact on game revenues, and therefore are receiving more and more interest [28, 39]. The systems we propose in this study can be seen as such an analysis tool to help the governance of an online social game.

To name a few, Xie et al. used decision trees to predict the level of engagement of players by using past data from other players [40]. Bauckhage et al. recently provided an overview of recent applications of clustering on behavioral data in games [41]. Hadiji et al. used behavioral data like playtime, session length and intervals to predict churn (i.e. players that no longer play the game) in five different free-to-play games and reported a high accuracy [42]. Similarly, [43] studied clustering of behavioral patterns for generation of user profiles and evaluated several unsupervised techniques on a large dataset of *World of Warcraft (WoW)* players, spanning five years of gameplay interval. The question of how to apply AI techniques to improve game moderation (rather than for implementing revenue increasing game changes) is still open. Analytics on digital games and analysis of big data are successfully used for several purposes in many games. There have been several studies on game data in academic literature.

Game analytics have been used previously for detecting different types of players. For instance [44] 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. [45] conducted a survey study on 1040 *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 *Extroverts* 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, [46] 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)* [47], 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 [48].

In [49], Miller and Crowcroft studied movements of players within the game world of *WoW*, identifying certain patterns and player behaviours in order to help game developers. Lewis and Wardrip-Fruin on the other hand, focused on analyzing player progress and tried to evaluate the balance and difficulty of *WoW* when played with different classes of avatars [50]. Similarly, Ashton and Verbugge studied game pace in *WoW* to assess difficulty[51].

In another study, [2] 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 [52].

In *Halo Reach*, developers gathered data from more than 3 million users in a period of 7 months and analyzed their gaming pattern, achieved skills in order to evaluate the effect of breaks between game sessions [53]. Using the same data source, Mason and Clauset and reported that players with more friends in the game perform better both individually and as a teamc̃itemason2013friends.

In *Project Gotham Racing 4*, the analysis of data from around 3 million players

to realize the distinction between long-term and short-term players, difference between multi and single player gaming patterns and utilization of various game options [54].

Earlier in 2009, Weber and Mateas studied *Starcraft* using replay data of 5.000 players in order to predict player strategies using machine learning techniques. Later, their results are enhanced the AI bot that can play competitively against human players [55]. Later, Weber et. al. studied analytics data of *Madden NFL 11* to enhance game retention, a common metric to evaluate the success of a game in the industry [56]. More recently, Yan et. al. analyzed data from 3.000 players of *Star Craft 2*, to evaluate the use of a game feature called "control groups" and reported how this feature alone correlates with the success of players in the game [57].

There are also several studies that present surveys and exhaustive overviews of literature on big data and game analytics on player behaviour, analysis and visualization [58, 59, 28].

3. Okey Game and the COPA Database

In this section, we introduce the dataset that we used in our studies. We start by giving an overview of the traditional game of *Okey* and the online versions in Section 3.1. Next, we focus on a specific version of the game that we gathered our data from in Section 3.2. We explain the data collection and annotation efforts in Section 3.3 and preliminary data processing made in order to extract relevant features in Section 3.4. We complete the section by explaining what features are available to us and how we categorize them in detail in Section 3.5.

3.1. An Online Social Game: Okey

Like in many countries, traditional Turkish games also have seen their online counterparts hitting the markets. A very well-known example of such a game is *Okey*, probably of Chinese origin, but adopted in Turkey and played socially for more than a hundred years. Okey is a part of the "Kıraathane" (i.e. the coffeehouse, but the word comes from "reading house") culture in Turkey; the coffehouses are social gathering places for men, who spend long hours there to drink tea and coffee, smoke cigarettes and hookah, read and discuss newspapers, and play card games, backgammon, and Okey. These coffeehouses have existed in Turkey since the sixteenth century [60].

The Okey game is played with four players, seated around a table. Players use wooden boards and a set of numbered tiles (See Fig. 3.1). We describe here a somewhat simplified set of rules to give the reader an idea. The tiles of the game resemble cards, and multiple independent rounds make up the game, which strongly resembles Rummy. Players draw and exchange tiles to finish the round first with a matching set of 14 stones. One of the stones is opened at the beginning of the round. The stone that follows it in value is called the Okey stone, and acts as a joker. Every player starts with a fixed number of points. Each individual round of the game has a single winner and every other player loses 2 points from their total points. The player in order takes one stone and then discards one stone on the table, which is seen by all players. The

Figure 3.1. Okey players in a coffeehouse.

player either draws a stone from a closed pile, or takes the last discarded stone. This is iterated until a player finishes the game, and the point totals are updated. Each game has a winner and betting is a common aspect of the game. The game mechanics require experience, attention, talent, and some luck, as in most games.

Okey is one of the first social games in Turkey to be carried on to the digital platform. Because of the requirement that there should be four players to play, bot programs are used to substitute for players who leave the online table. Since the social aspect is the primary aspect of the game, online Okey sites have also offered means of player interaction, typically in form of a chat window alongside the main game window. Depending on the experience of the players, each Okey game lasts around 5 to 10 minutes. As a casual game, generally during the turn waiting period, players use the chat area to socialize and enjoy their time.

While a simple game, digital versions of Okey attract a very large audience. There are several Okey games developed by various Turkish game companies and even individuals that are available freely online. According to the independent statistics service AppMtr, the five most popular online Okey games on the Facebook platform

have a total of 6.500.000 monthly active unique users as of December 2014[22] .

Figure 3.2. Online version of Okey. Chat area is in bottom left. Players have unique and customizable avatars. Player on the left side is a bot. Players and visitors in table are listed on the bottom middle area and can interact with chat and gifting. Player at the top has a gift (drink) next to her avatar.

In the online version of the game, the leading player sets up a table in order to start a game, and other players join in. Once the game is played out, the players may disband and join other tables, or if the table is sufficiently entertaining, they may continue playing together, with a bot replacing any resigned player. Each game has a single winner and betting with virtual currency is a common aspect of the game.

Online versions of Okey typically offer a chat window next to the main game window, as the social aspect of the game is very prominent. Often during the turn

[22] AppMtr - Facebook App Usage Metrics Tracking, available at `http://www.appmtr.com/search/?q=okey`. [Online; accessed 12-December-2014]

waiting period, players use the chat area to socialize and enjoy their time. Players can also establish opt-in friendships by sending in-game friendship requests to other players. Once the virtual friendship is established, both players can send each other offline messages within the game even if they are not in a game table. The offline messages arrive at players' virtual message inbox when a player is not online when messages are received.

These communication channels sometimes see abusive or offending message exchange between parties. In such cases, players can submit complaints about other players. Each player has a link to their player profile that shows their avatar, nickname and game statistics. There is also a complaint submission option in the profile page that lets other players to enter some details of their complaints about the player and submit to game moderators.

If a player abuses or offends one or more players, these need to be dealt with quickly and decisively in order to keep the ambiance of the game intact. Possible actions taken by game moderators include sending warning message to the player, or banning the player permanently or for a certain a period of time, depending on the severity of the abusive act.

Players have means to submit complaints about other players. However, analyzing these cases and deciding on a verdict requires a thorough investigation of both parties' profiles, game histories and chat logs in game tables they shared. For game moderators, dealing with these incidents requires utmost attention and is a very time consuming task. The lack of robust player profiling and analysis tools to help human moderators dealing with such complaints is a general problem for all social gaming industry. The stress on the social dimension makes the Okey game particularly adequate for adapting real-time human behavior analysis approaches.

3.2. CCSoft Okey Player Abuse (COPA) Database

In this study, we collected, annotated, and used the proprietary CCSoft Okey Player Abuse (COPA) Database, consisting of player demographics, statistics, game records, interactions and complaints. The database is acquired from a commercial Okey game over a six months period, and incorporates roughly 100.000 unique players who at least played the game once. Around 30.000 of these players played the game at least five times.

All the player identification information is deleted prior to our analysis to anonymize the data and to protect players' privacy. In the mentioned period, a total of 800.000 Okey games were recorded along with the player interactions in the chat area. The total number of game table logs generated by the players during the period exceed 520.000 items.

3.3. Data Collection and Annotation

COPA set contains raw SQL exports of several relational database tables. The information related to a specific user is spread among several tables. We have developed a software to join all the tables and extract the features for each user and store in a more convenient format for further studies.

In addition to SQL database entries stored about the users, COPA also includes individual chat logs for every table created in the game. Each table log is stored as a JSON formatted text file containing pieces of information that represent the actions that happened in that table. In addition to individual chat entries from the players in that table, the log also contains actions such as game start and end, player entering and leaving, use of offensive language caught and prevented to appear on other player's screens, individual game options, along with timestamps for each of these actions. A sample excerpt from a table log file is shown in Figure 3.3. Note that individual player identities are hidden by the system and replaced by encoded texts automatically to prevent leaking of personal privacy.

```
[{"who":"KQqo9-urh0CXdd278c3MGgZ-","at":"2012-09-01T12:56:26.5864398Z","act":"table.create","obj":"{\"CanShowFaceup\":false
,\"CanChat\":true,\"IsPartnerOn\":true,\"Name\":\"Masa 31\",\"StartScore\":4,\"Seed\":null,\"TimeoutDuration\":15,\"Bet\"
:5000}"},
{"who":"KQqo9-urh0CXdd278c3MGgZ-","at":"2012-09-01T12:56:26.7739638Z","act":"table.sit","obj":"0"},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:57:04.0749365Z","act":"table.join","obj":"88.228.251.124:49839"},
{"who":"server","at":"2012-09-01T12:57:04.0749365Z","act":"chat","obj":"menu13 masaya girdi."},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:57:11.2630345Z","act":"table.sit","obj":"2"},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:57:32.2815838Z","act":"table.leave","obj":""},
{"who":"server","at":"2012-09-01T12:57:32.2815838Z","act":"chat","obj":"menu13 masadan ayrıldı."},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:58:04.4279065Z","act":"table.join","obj":"88.228.251.124:49839"},
{"who":"server","at":"2012-09-01T12:58:04.4279065Z","act":"chat","obj":"menu13 masaya girdi."},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:58:04.5998112Z","act":"table.sit","obj":"2"},
{"who":"KQqo9-urh0CXdd278c3MGgZ-","at":"2012-09-01T12:58:20.4932123Z","act":"chat","obj":" hg"},
{"who":"lCZAknmNFk6Ir5vI_5bgEg8a","at":"2012-09-01T12:58:23.4000017Z","act":"table.join","obj":"95.10.146.127:49817"},
{"who":"server","at":"2012-09-01T12:58:23.4000017Z","act":"chat","obj":"brk1982 masaya girdi."},
{"who":"lCZAknmNFk6Ir5vI_5bgEg8a","at":"2012-09-01T12:58:23.5719086Z","act":"table.sit","obj":"1"},
{"who":"KQqo9-urh0CXdd278c3MGgZ-","at":"2012-09-01T12:58:34.1676252Z","act":"chat","obj":" h.gdiniz"},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:58:34.8708897Z","act":"chat","obj":" h.b."},
{"who":"lCZAknmNFk6Ir5vI_5bgEg8a","at":"2012-09-01T12:59:10.8464206Z","act":"chat","obj":" hb slmlar"},
{"who":"KQqo9-urh0CXdd278c3MGgZ-","at":"2012-09-01T12:59:25.2083222Z","act":"chat","obj":" a.s"},
{"who":"vjLEX8a9xE20MRt0YH9uPwOJ","at":"2012-09-01T12:59:38.867438Z","act":"swear","obj":" eş lı mı oyun"},
{"who":"KebxHtwyYkWd7apSxSAWggyd","at":"2012-09-01T12:59:44.9154353Z","act":"table.join","obj":"88.233.67.62:55254"},
{"who":"KebxHtwyYkWd7apSxSAWggyd","at":"2012-09-01T12:59:45.1029701Z","act":"table.sit","obj":"3"},
{"who":"server","at":"2012-09-01T13:03:41.2493934Z","act":"chat","obj":"huba2005 eli kazandı."},
{"who":"KebxHtwyYkWd7apSxSAWggyd","at":"2012-09-01T13:10:24.487483Z","act":"table.leave","obj":""},
{"who":"Çeker Atar","at":"2012-09-01T13:10:24.800009Z","act":"bot.fill","obj":"3"},
{"who":"server","at":"2012-09-01T13:10:24.800009Z","act":"chat","obj":"Ahmet A. masadan ayrıldı."},
{"who":"lCZAknmNFk6Ir5vI_5bgEg8a","at":"2012-09-01T13:10:27.3314696Z","act":"chat","obj":" puan nasıl alınıyo"},
{"who":"KQqo9-urh0CXdd278c3MGgZ-","at":"2012-09-01T13:10:37.1291018Z","act":"chat","obj":" kazandıkça"},
{"who":"server","at":"2012-09-01T13:12:08.4631424Z","act":"chat","obj":"menu13 eli kazandı."},
{"who":"server","at":"2012-09-01T13:15:50.9590234Z","act":"chat","obj":"brk1982 eli kazandı."},
{"who":"server","at":"2012-09-01T13:18:21.8455548Z","act":"chat","obj":"menu13 eli kazandı."},
{"who":"QRn7L81DakKjsNSY24eFWgZX","at":"2012-09-01T13:20:21.9913803Z","act":"table.join","obj":"94.123.131.232:58449"},
```


Figure 3.3. Sample table log contents.

From the table logs, one can also retrieve chat related information such as how frequent a player initiates a conversation, how responsive and talkative a player is, correlation between winning or losing and chatting by analysing the timestamps of each action in a log. The administrative interface of the actual game environment has a screen to observe each table log so that moderators can study and work out whether a complaint submitted in a table has valid reasons and requires further administrative action.

For our studies, a software that parses and interprets these table logs is implemented. The software extracts all the relevant information for each complaint and involved user in the form of a feature vector.

For each player in the database, we process this set of table logs, filter the ones that involve the related players, and extract the number of chat and bad language attempts to use for player profiling. Within the data collection time frame, 1.066 of the 5.000 submitted player complaints were manually labeled by game moderators, as being abusive or not abusive. Note that game moderators initially only read the contents of the complaints, but a full annotation requires checking further evidence. These complaints were reported by 726 unique players and a total of 689 players were

accused. Some of the abusive players involved in more than one complaint. The remainder of the complaints are annotated as being not abusive. Our annotations include a subjective judgment of the severity of each abusive case, denoted on a Likert scale (1 being the least severe, and 5 being the most severe).

In order to create training and testing sample sets for complaint and player analysis, we have to manually annotate complaints that are related to genuine abusive action or harassment. To achieve this, labeled complaints were thoroughly analysed by inspecting table logs and private messages between involved parties (i.e. accuser and accused players). For each complaint, this manual process take around five minutes on average. Therefore, around 5.000 minutes of manual effort is spent during this procedure in total. As a result, 240 players involved in these complaints were found out to have actually harassed others to different degrees. The remainder of the complaints are annotated as being not abusive by the moderators.

During the evaluation of complaints, we also annotated complaints with additional labels according to the following content:

- Obscenity and use of swear words
- Abuse against opposite sex
- Mutual fault with both parties abusing each other
- Sarcasm and/or mild humiliating language
- Insulting
- Serious nervosity and tension

We also studied what types of abusive cases exist in the dataset and try to categorize them. Table 3.1 lists the types of complaints that we have encountered, and their relative frequencies in overall complaints.

Along with abuse types, we also ranked the severity of the abusive actions on a scale of 1 to 5. For instance sarcasm is a very mild type of abuse, but it can lead to more serious aggression, and subsequently included in the annotations. A mild dispute

ending up with some foul language has a severity of 1, while a targeted, repeated abusive act containing racist or sexist terms will have a severity of 5. Severity ranks turned out to show a flat distribution in our complaints database, where each cluster with a specific severity rank had around 20% share of the total number of complaints. Similarly, we found that severity is distributed evenly over the abuse types listed in Table 3.1.

Abuse Type	**Description**	**Occurrence (%)**
Obscenity	Use of bad language	78.29%
Sexual Harassment	Mostly against opposite sex	62.30%
Insult	Humiliating language use	55.79%
Sarcasm	Indirect, sarcastic comments	25.53%
Mutual	Two parties abusing each other	31.91%

Table 3.1. Different types of abuse and their distribution in the COPA dataset. Complaints may have more than one type of abuse involved.

3.4. Preliminary Analysis

Among over a thousand actual player complaints which were processed during our database preparation studies, only around half of the investigated complaints were determined to contain really abusive acts and required administrative action from game moderators. There are several reasons as to why the ratio of genuine complaints is so low.

In some cases, victims submit complaints for a player who was not actually involved in any abusive behavior at all by mere mistake. These are due to shortcomings of the complaint submission interface. As we noted earlier, in order to submit a complaint related to another player, one has to open that player's profile page, click a link and fill in the details of complaint. We have encountered several players using complaint interface to send actual personal messages to these players although the link clearly states it is for submitting complaints for that person (in Turkish, the link reads

"Şikayet Et").

In several cases, there are complaints submitted but without any details. Although these cases may be genuine and need to be dealt, it is very difficult to guess what the complaint is about, therefore generally ignored by moderators and annotators.

Another common case for false complaints are actually submitted by abusive players. When they are involved in heated disputes and violate the rules of conduct, although they are the guilty party, they submit complaint about the victim claiming that they were harassed first. This is a very common pattern, which shows us that the players think the game moderators do not commit to the analysis of these cases but give their verdict by looking at players' game statistics alone. One other frequent reason for false complaints happen when a player loses a game and feels cheated. In order to inflict damage to the opponent player, they claim abusive acts occurred.

Finally, a host of complaints arise from the cultural diversity in the player background: A friendly chat sentence, or a gifting gesture can easily be misunderstood or interpreted as sexual harassment.

It is our premise that an automatic system can deal with such behavior patterns, provided that sufficient examples help in the training of the system. Such an automated system can analyse and prioritize the complaints for human moderators, saving costly human resources.

The complaints are not necessarily reported during a game, a player can choose to report an abuse afterwards when in lobby or in another game. Therefore, we cannot simply pinpoint each game log associated with the complaint. We select a set of candidate of game logs that contain both the reporter and the accused, with a timestamp of game start prior to the submission of the complaint. With this filter, we isolated 1.222 game logs for further analysis.

The game incorporates a basic bad language detection system that blocks a set

of black-listed words and their variants. When a word in the black list is intercepted, the system blocks it and informs the offending player. Very often, warned players will rework the offending sentence to bypass the filter and they will usually manage to submit an abusive sentence by tweaking the letters and words. In addition, the filter has been reported to be only moderately effective, as it generates a large number of false positives when used with a large black list vocabulary. Therefore, this filter is used with a small vocabulary and bad language interception is not counted as direct evidence of actual abuse.

In addition to chat, players have several other means to socialize during the game. They can send virtual friendship invitations to each other and establish virtual friendships, which enables them to send offline messages to one another and to locate a particular game they play when both parties are online. The game also take advantage of Facebook social integration tools, in which players can invite their Facebook friends and establish in-game virtual friendships, and share their game achievements in their Facebook profile. In exchange of Facebook interaction, players are rewarded bonus virtual currency, which can be used for in game betting. Using a standard double currency system, players can also pay real money to buy virtual currency. With virtual currency, players can purchase virtual gifts in game tables and send them to other players in the game. These means help people to enjoy their turn waiting time, increase the engagement and decrease the virtual currency inflation. One hypothesis we entertain is that players with high social activities may tend to offend others more, as well as being offended easily. We also argue that paying players would take the game environment more seriously and therefore will not offend others easily, but report abuses and other offensive behaviour. Therefore, the amount of social interaction and real life payment statistics are also included in our feature vector.

3.5. Features

For each player, we extract information that constitutes a player's profile. In order to form a detailed profile that can help moderators during evaluation of player complaints, we analyse table chat logs and players' game performance and retrieve a

number of features related to the player from the game database as we explained earlier. In addition to accuser and suspected players' profile data, we augment the feature set with information related to communications between these two parties involved in the complaint. The list of features used and their brief descriptions are given in Table 3.2.

[28] classify game user metrics into three categories as *gameplay metrics* (variables related to the behavior of the player in the game), *customer metrics* (all aspects of the user as a customer), and *community metrics* (metrics that relate to the community and social interactions). We classify our features according to this taxonomy, as summarized in Table 3.3.

The first set is comprised of *gameplay* related features, which signify a player's performance in the game. *Customer* related features on the other hand summarize the information related to the player, disregarding the game specific elements. Lastly, *community* features contain data about the social interactions afforded by the game environment. While these differ from game to game, all social online games provide a set of similar features that can be substituted in the analysis of a different game.

We would like to emphasize here that the *customer* and *community* features are not game-specific. However, *gameplay* related features incorporate to a certain degree elements which are specific to the Okey game itself, such as ratings and counts of incomplete games. These features may not have a counterpart in other games, or their significance may be different. We assume that for each game, this part of the feature set will contain specific and potentially relevant game indicators.

In following sections, we describe all features in three feature sets we used in our experiments individually. We also report some statistics related to these features to make them more tangible. The list of used features and their category assignments are given in Table 3.2.

3.5.1. Gameplay Features

Games Played: The first gameplay feature we consider is the number of played games over the entire history of a player. A high number of games played indicates an experienced and engaged player, who knows about the social rules and conventions of the game community. If we look at the 100.000 players in our entire game database, they have each played 33 games on the average.

Number of Wins: A high rate of wins per game played reflects the player's performance related to the specific game, and therefore it is included in our feature set. A player with high number of wins is assumed to have a better understanding of the game community, and would be reluctant to engage in aggressive and abusive acts in order to preserve personal reputation and overall game score. The perceived success is assumed to be correlated with certain gaming behaviors like seeking out strong opponents and reputation building.

Incomplete Games: Players can leave a game before it is finished properly. This may happen due to valid reasons, such as losing connection to the game servers, or having a real life interruption, like a phone call. Other times, a player can understand that he or she will lose the game eventually, and leave the game before it ends. In this case, the game system replaces the player's seat with a bot, far less talented than an average player. Since the game is played by four players, the less talented bot gives an advantage to the players sitting next to the bot, and disrupts the balance of the game. While in some games surrendering can be a sign of respect, for the Okey community, leaving the game is considered as being disrespectful to fellow players. The cause of someone leaving the game table can also be attributed to an incident happening in the table, for instance due to a dispute or harassment. These cases are obviously very much related to our analysis. Repeatedly leaving a game before the end can be a sign of bad or rude behavior.

Rating: The game has a rating system, similar to the well known Elo rating system commonly used in chess and similar games [61]. Basically, a rating system

evaluates each game played by a player, taking into account how good the opponents are. It is used to eliminate lucky wins or losses, and to favor wins against players with comparable skills. Unlike Elo, Okey's rating algorithm can cope with games played by more than two players, and updates effective ratings for each player at the end of a given game. We use rating as a gameplay feature, since it signifies the competence of the player and there might be a correlation between aggression and abusive behavior among the players of varying skills. Although in this study we do not evaluate the contribution of individual features to the overall results, we believe player rating measures competitiveness, which is likely to be a factor for the player's behavior.

3.5.2. Customer Features

Gender: When the players first register for the game, they choose the gender of their in-game avatar. This is the only available demographic information related to the players, but also very essential in assessing the submitted abuse complaints. Since the players often connect their gaming account to their Facebook accounts, we observe that players mostly choose avatars consistent with their real gender[23] . We also observe that, 93% of our players marked as genuine abusers use male avatars. Note that only 70% of the overall player population is male in our database. While the clientele of real coffeehouses is predominantly male (except around some universities, where both male and female students visit them), Okey is enjoyed by both male and female players, and is popular as a home entertainment.

Virtual Currency Purchases: Using a double currency system commonly implemented on most social games, players can pay real money to buy virtual currency within the game. These virtual currency can later be used for in game betting, avatar customizations and for other similar purposes.

[23]The percentage of user-misclassified accounts is assumed to be low. According to the Quarterly Report Pursuant to Section to 13 or 15(d) of the Securities Exchange Act of 1934, for the quarterly period ended June 30, 2013, filed to the Securities and Exchange Commission (SEC), Facebook estimates that user-misclassified accounts may approximately represent 1.3% of all Facebook accounts on the average, with Turkey stated as probably having a higher percentage.

Purchasing behavior is a promising feature to contribute to a player's profile, as it shows the amount of real assets spent for the sustenance of the game. Our gaming database contains a small number of purchases: around 2% of the players did purchases, for both abusive and non-abusive groups of players. Nonetheless, we incorporate number of purchases for each player in our feature set.

Daily Logins: Number of daily logins to the game is high for experienced and loyal players. Players who spent more time in the game tend to fit in with the social rules of the game better and tend to get less complaints compared to new players. On the other hand, short-lived accounts may have been created by existing abusive players who are already banned or tagged as abusers in the community and avoided by other players. In the database, 15% of the players marked as abusive logged in to the game for only one day. Around 41% of them appeared in game for less than 5 days. For the non-abusive players, however, only 12% have a login count less than 5 days. In light of the evidence from the dataset, we argue that daily login counts can be used as a feature to distinguish abusive players.

Tables Entered: Similar to daily login count, the number of tables entered by a player signifies the time spent in the game. It also hints at the openness of a player to meet and play games with new people.

3.5.3. Community Features

Friends: In addition to chat, players have several other means to socialize during the game. They can send friendship invitations to each other and establish virtual friendships, which enables them to send offline messages to one another and to locate a particular game in progress when both parties are online. One can argue that a player with a high number of friends can fit in the environment and should not cause any complaints. On the other hand, friendship enables players to send private messages to each other, which can become a medium of abuse when used with bad intentions. Novice players can be targeted with such harassment, as they are more prone to accepting new friendship requests, since they have not grasped the social dynamics of

the game environment yet.

Friendship Requests: During our preliminary analysis of complaints and abuses, we realized that repeated sending of friend invitations itself can be perceived as a form of harassment to some players. Especially female players (as indicated by their avatar) submitted several complaints about players who insisted on sending invitations. While players can completely ignore these invitations, they are probably causing nuisance and perhaps further disputes. We decided to add the number of friendship requests to our feature set to explore its contribution.

Private Messages: When players create in game friendships, they can send offline text messages to each other. These messages are stored in their private inbox, similar to an email system. Private messaging is a means to socialize players who do not know each other in real life. The amount of such messages sent and received are used as features.

Social Rewards: The game takes advantage of Facebook social integration tools, in which players can invite their Facebook friends, establish in-game virtual friendships, and share their game achievements in their Facebook profile. In exchange of Facebook interaction, players are rewarded bonus virtual currency, which can be used for in game betting. We use the amount of social rewards obtained in this manner as a feature.

Gifts Purchased: With virtual currency, players can purchase virtual gifts in game tables and send them to other players in the game. This is a typical mechanism to help players pass the time otherwise spent waiting for other players to finish their turns. The number of purchased gifts is used as a feature.

Avatar Items: Most social online games provide ways to customize the look of the player. The players can harness these features to select an appearance that reflects their own personality, preferences, or to make a statement. In the Okey game, a player can customize his/her own avatar, change the facial look and purchase clothes and form a virtual wardrobe. Players can use either real money or in-game currency to buy such

items. Some players tend to ignore this feature completely and use the default avatar created automatically at the start of the game. In some online environments (like the Second Life), using a default avatar is frowned upon, and is bad form. In Okey we do not notice such an attitude. Nonetheless, some players compose a full wardrobe to create a good looking avatar.

It is possible that avatar customization is a means of socializing and a proof of engagement in the game. Furthermore, people who go to lengths at customization may receive differential treatment from their fellow players, for instance provoking more aggression, or sexual harassment. In this study, we only looked at the number of avatar items as a feature.

Chat Entries: During the gameplay, players who are playing the game and players visiting the table without playing have a common chat area. In our analysis, we process all table logs of a given player and process all the related chat entries. We store the number of chat sentences entered to designate talkative players. Around 10% of the players in the database have no chat records, meaning that they are not interested in the socialization aspect of the game. Since the act of abuse mostly occurs in the chat environment, all abusive players have some chat records, as can be expected.

Chat Word Ratios: In addition to chat entries, for each sentence entered, we also count the number of words used during chat. One observation from the data is that, most players use chat area for short utterances, longer sentences tend to be a better predictor of a real social interaction. In order to detect long sentences, the total number of words used in chat area for every player is divided by their total chat input lines. Average word count for abusive players is 1.8 and the abusive player with the greatest chat word count has uttered 3.9 words per entry on the average. When we look at the same values in the full database with 100.000 players, the average word count is 1.1 and the most talkative player used 22.2 words per entry. Only 5% of the players have not used the chat feature at all, which illustrates the importance of the social aspect in this game.

Silence Before/After Chat Entries: Although chat entries and number of words are good estimators of socially active players, they do not directly indicate social interaction. For assessing social interactions, we analyze non-verbal cues [62]. We measure the *silence* before and after each chat entry, and normalize these values according to total length of actual chat entries for each player. A long pause after a chat entry can imply the end of a communication, while a long silence before an entry can show initiation of a possible conversation. In the literature, a similar analysis of silences between conversation segments was recently presented in the audio domain by [63], where the authors used a corpus of formal meetings, and used the silences effectively to detect social verticality constructs, such as dominance and leadership.

Bad Language Attempts: The game incorporates a basic bad language detection system that blocks a set of black-listed words and their variants. When a word in the black list is intercepted, the system blocks it and informs the offending player. Very often, warned players will rework the offending sentence to bypass the filter and they will usually manage to submit an abusive sentence by tweaking the letters and words. If the filter is used with a large black list vocabulary, it can be only moderately effective, as it will generate a large number of false positives. Therefore, this filter is generally used with a small vocabulary. Bad language interception is not counted as a direct evidence of abusive behavior, but it is indicative of intent, and as such, used as a feature. Offensive language directed at self (e.g. 'sh.t, made a mistake!') may not count as verbal abuse, but 'sh.t, you're a terrible player!' would count. Context is relevant for determination of toxicity, but in the present system it is largely ignored.

Feature	Description
Games Played	Number of games played by player.
Wins	Number of wins of player.
Incomplete Games	Number of games player left before game ends.
Rating	ELO-like rating of the player, calculated in a multiplayer setting.
Gender	Gender of the player as they declared during player creation initially.
Credit Purchases	Player's total number of virtual currency purchases using real money.
Daily Logins	Number of logged in days to the game.
Tables Entered	Number of tables joined (for both playing and watching others play).
Friends	Number of in-game friendships made. The game offers invitation based friendship mechanism.
Friendship Requests	Number of in-game friendship requests sent. Some may have been unaccepted.
Private Messages (PM)	Number of private (offline) messages sent to in-game friends, delivered at log in.
Social Rewards	Number of social rewards earned by actions such as Facebook shares and likes.
Gifts Purchased	Number of virtual in-game gifts (visuals placed next to player seats) purchased in game tables.
Avatar Items	Number of items purchased to customize avatar used throughout game tables and rooms.
Chat Entries	Number of chat utterances made in all tables entered.
Words in Chat	Number of all words uttered in all chat sessions.
Silence Before Chat	Average time passed in table before each chat utterance of player.
Silence After Chat	Average time passed in table after each chat utterance of player.
Bad Language Attempts	Number of utterances that are detected and prevented by the system as foul language.
Accuser Chats	Number of chat utterances by the accuser in last 3 shared tables with suspect.
Suspect Chats	Number of chat utterances by the suspect in last 3 shared tables with accuser.
Other Chats	Number of chat utterances by other players (excluding accuser and suspect) in last 3 shared tables.
Accuser Words	Number of words uttered by the accuser in last 3 shared tables with suspect.
Suspect Words	Number of words uttered by the suspect in last 3 shared tables with accuser.

Table 3.2. Features in the data set and their descriptions.

Category	**Feature**
Gameplay	Games Played
	Number of Wins
	Incomplete Games
	Rating
Customer	Gender
	Virtual Currency Purchases
	Daily Logins
	Tables Entered
Community	Friends
	Friendship Requests
	Private Messages
	Social Rewards
	Gifts Purchased
	Avatar Items
	Chat Entries
	Chat Word Ratios
	Silence Before Chat
	Silence After Chat
	Bad Language Attempts

Table 3.3. Features in the data set and their categories.

4. Proposed Methodology

In this study, we propose two separate systems for two related problems. First, we propose a system to classify abusive players who are reported in game complaints. For this task, we mainly rely on the data specifically related to a single player who is reported to be abusive in the complaint. Second, to automatically analyse and rank player complaints, we develop a system that incorporates information from both players who are involved in the dispute; the suspected player and the victim player. The latter system is built upon the former one, by critically enhancing the feature vector with user features from the victimized player. In both systems, we incorporate the same COPA dataset.

We also studied further verbal communication in our chat data. In a multi-party chat environment, using an automated procedure, we estimate whether chat content contains verbal aggression by affect analysis. Finally, we address another common problem in dealing with abusive actions in a game environment where abusive users generate several user account. Using the same chat data, we try to automatically recognize authorship and propose a system that will cluster such users.

In this section, we first briefly overview our general approach to the problem of classifying abusive players and genuine complaints in Section 4.1. Next, we explain the preprocessing and evaluation schema that we incorporate for these classification systems in Section 4.2. Next, we briefly layout the flow of our methodologies on abusive player classification and complaint classification in Sections 4.3 and 4.4.

4.1. General Approach

In both systems, we use machine learning to automatically classify positive samples. From the dataset, a training and benchmarking set is generated, where player complaints are manually labeled as ‘abusive’ or ‘offending’ by human moderators as detailed in Section 3. For the scope of our study, a single moderator is assigned for

annotating complaints. Multiple annotators would certainly increase the quality of annotations, but at the cost of doubling or tripling the annotation expenses. The information of players involved in these complaints is extracted from the central game database as also described previously.

In order to have balanced priors for abusive and non-abusive players, we have randomly selected a group of 240 players from the database who have played at least 5 games and were never involved in any complaints. Together with the abusive players, these 480 players form our player database to be used in training and testing.

We train our classifiers on a portion of labelled data, and then report results on novel samples. To make full use of available annotations, we used a 10-fold cross-validation scheme. As a preprocessing step, we standardize all features individually to have zero mean and unit normal [64]. For the standardization step, we incorporate the whole set of players who have played at least 5 or more games (30.000 players). We observed that inclusion of this constraint improved the hit rate, as well as sensitivity and precision (3-5% in each score type), compared to standardization with 480 samples. Using the additional players in classifier training as well, however, causes a severe data imbalance problem, and deteriorates the results.

4.2. Preprocessing and Evaluation

For each feature value x, the standardized value x' is calculated as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{4.1}$$

where μ is the mean and σ is the standard deviation of the population.

More formally, let f_i represents ith feature for a category k with N distinct features. We calculate F_k, the scalar value which represents the category in our new

feature vector, as follows;

$$F_k = \frac{1}{N}\sum_{i=1}^{N}(f_i * w_i) \tag{4.2}$$

In this study, we ignore the weights w_i and use 1 for each.

In order to evaluate our results, we count the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) to calculate sensitivity, precision and specificity as well as accuracy. In information retrieval, these terms are defined as follows:

$$precision = \frac{TP}{FP + TP} \tag{4.3}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{4.4}$$

$$specificity = \frac{TN}{TN + FP} \tag{4.5}$$

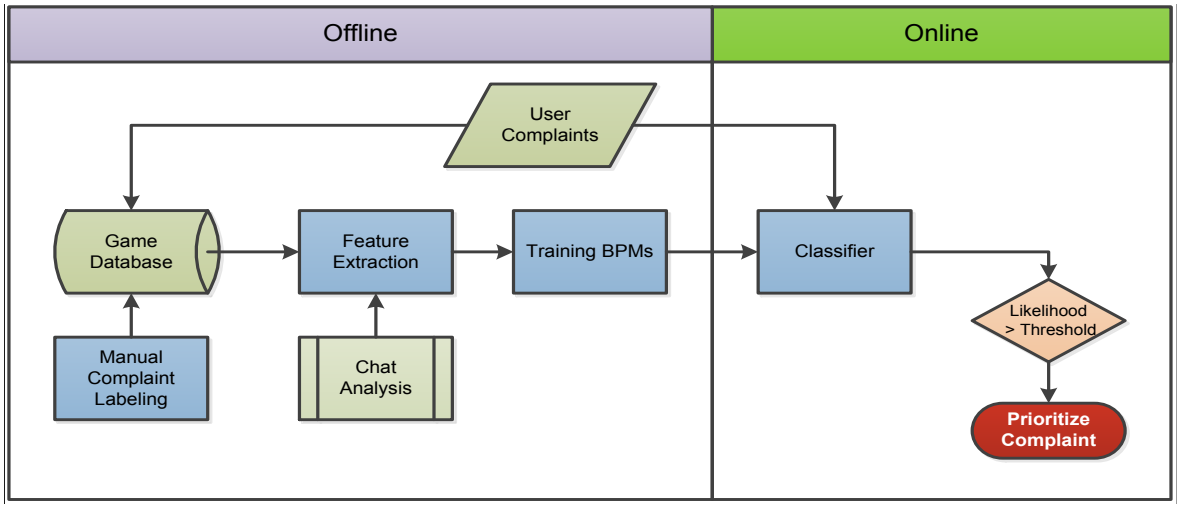


Figure 4.1. The basic flow of the proposed system for offline training and online classification for abusive players.

$$accuracy = \frac{TP + TN}{TP + TN + FP + TP} \tag{4.6}$$

In general, one will try to set a confidence threshold that results in high accuracy, precision and specificity, so that human moderators can prioritize such complaints. High accuracy and precision means that the system catches truly abusive players and filters out questionable complaints, whereas high specificity means that false accusations are rejected with a high probability.

4.3. Abusive Player Classification

In order to decide whether a player falls into the offender category or not, we train a supervised binary classifier, where the two classes stand for *genuine offender* and *not an offender*. This procedure is done in an offline manner. When a new complaint arrives during the game, the proposed system evaluates the accused player's profile, giving a likelihood of the player to be a genuine offender. Fig. 4.1 gives an overview of the basic flow of the system.

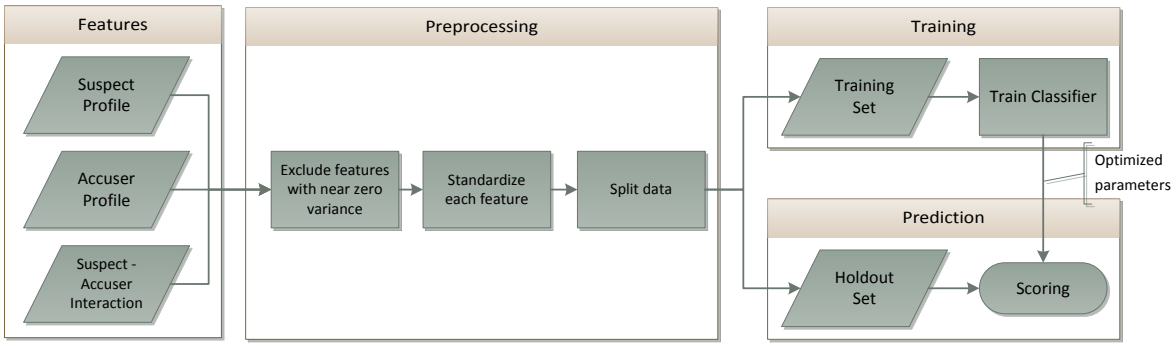


Figure 4.2. The basic flow of the complaint classification system for offline training and online classification.

For this study, we used Bayesian Point Machines (BPM) as our classifier. The BPM algorithm is explained in Section 4.7.4.

The BPM classifier outputs the likelihood of a player to be in the offender group. Since it is a binary classifier, for each sample, a likelihood value greater than a certain confidence threshold results in the classification of an offender. This threshold parameter trades-off sensitivity and specificity.

We perform experiments with several feature sets. In the first set, we treat each feature as being equally important, and form a single feature vector of 19 dimensions. We call this feature set the *combined* feature set. We also form separate feature vectors for the categories we laid out in Section 3, namely *gameplay*, *customer*, and *community* features. Separate classifiers are trained for each group of features.

4.4. Complaint Classification

For classifying complaints, we again train a supervised binary classifier, where the two classes stand for *complaint with genuine abusive acts* and *not a genuine complaint*. Similarly, training is done in an offline manner. When a new complaint arrives during the game, the proposed system evaluates the both offender and accused player's profile, analyzes communication involved both parties and score a likelihood for the complaint under test. Fig. 4.2 gives an overview of the basic flow of the system.

For this study, we used Gradient Boosting Machine (GBM) as our classifier. The GBM algorithm is explained in Section 4.7.5.

We report results with two different feature set configurations. The first configuration involves only suspected player's profile as input. This setting helps us to compare our results with our previous study [18], in which we also used the COPA dataset, but classified abusive players instead of classifying complaints. This configuration will establish a baseline to measure the impact of augmenting the dataset with accuser profile and features derived from the recent communication between accuser and suspected players.

In the second configuration, we present results with the entire feature set detailed in Section 3, including suspect and accuser profiles and their interactions.

As noted earlier, the whole data set contains 907 complaints. For the baseline experiment, which uses only the suspected player's profile, the feature vector has 19 dimensions. For the second experiment, we incorporate both accuser and suspect player profile data and their recent relevant interactions. Using 19 features per player, and 5 features related to inter-player communication (see Table 3.2), we obtain a feature vector with 43 dimensions.

For each training and prediction configuration, we used 75% of the population for training a GBM and reserved the rest as a holdout (test) set. During training, we apply a 5-fold cross validation scheme for parameter estimation and fine tune the number of trees to be used for prediction to prevent overfitting. The holdout set is not used in this phase. We used a *Bernoulli* distribution to model the loss function, which is the suggested approach for binary classification. We used half of the samples without replacement for subsampling in learning iterations in order to introduce some randomness and to prevent overfitting.

4.5. Verbal Communication Analysis

In this section, we look at the problem of assessing affect in multi-party chat by automatic analysis of the chat text. Most languages do not have the rich text analysis tools developed for English language; we illustrate in this section how some of the tools originally developed for English can be adapted to a different language (i.e. Turkish), and assess the accuracy of the ensuing system. The proposed approach could be used as a stand alone module, or for complementing multimodal approaches to analyze computer mediated communication.

Text-based online communication, such as emails, text messages, or instant message chats, have come to constitute important methods for communication and collaborating among group members. This new age communication trends open novel areas of research as well such as affective analysis in these online communication platforms. There have been several studies on sensing the affective state using various physiological and behavioral signals [65], such as using facial expressions [66], text, speech, or cognitive and semantic cues [67]. However, a few of them have applied this information to online communication systems, such as online chat. This highlights the need for further research on analysis of affect in online chat environments.

Compared to other forms of affective content, text is believed to be relatively less challenging to analyze. However, unlike well-structured documents such as newspapers, books, and blogs, chat conversations are much more chaotic, and full of irregularities. On the other hand, they are also more loaded with affective content. Consequently, analysis of chat text for affect is far from trivial, but potentially rewarding.

We present here an affective analysis model for the Turkish language. The system was specifically designed for and tested with in-game multiparty chat logs, where the language is typically very flexible, irregular and emotive. We have collected and cleaned a large amount of chat data, and annotated them with syntactic and lexical semantic information. The second step was the adaptation of an affective lexicon from English to Turkish, and the creation of automated tools to analyze the chat corpus. Finally,

we created an affect analysis model, which uses 120 emoticons, 98 abbreviations, 50 interjections, 72 modifiers, and the affective lexicon that includes more than 15,000 words. This model was used to cope with abbreviated and informal chat language and to capture a wide range of affective features.

The section is organized as follows: Section 4.5.1 gives a brief explanation about the studies that have been proposed to investigate affective states and processes in online chat. Section 4.5.2 discusses the corpus, the annotation procedure, the assessment of annotation quality, and the creation of the affective lexicon for Turkish. Section 4.5.3 presents the affect analysis model. Section 4.5.4 describes the evaluation of the model, the results, and the limitations of the model. Section 4.5.5 concludes the section.

4.5.1. Related Work

Many researchers designed and implemented applications and intelligent user interfaces (such as email composers, online chat interfaces or instant messaging systems) with rich integrated emotion conveying engines. Sánchez et al. designed an instant messaging system called *Russkman* [68]. This system is enhanced with functionality that allows users to convey moods and emotions while interacting with other users. *iFeel_IM!* is another system that enables users to express emotions during online communication [69]. *CrystalChat* visualizes a user's social network by extracting user's chat log history and with the help of a graphical interface [70]. It presents patterns of the conversation length and emotional tone, based on the emoticons used. *EmotionChat* is designed as a chat platform between teachers and students in e-learners systems [71].

These applications allow users to express emotions, but there are other studies that sense affect automatically, for instance by including a tactile emotional interface for instant messenger chat [72], or by using physiological data and animated text with the help of a physiological sensor attached to the body [73]. In the absence of multimodal input, just the text is analyzed to predict emotions [74].

Current approaches for automatic emotional and affective content analysis from text generally include keyword spotting, lexical affinity, statistical natural language processing (NLP), learning based methods and commonsense-based approaches [75].

The most straightforward approach is keyword spotting, which identifies a set of keywords to construct a look-up table that contains keywords and their affective values. The basic limitations of this approach is that it is incapable of dealing with negation and with complex sentence structure. Building a rich lexicon is a very expensive task, and affect-conveying words only form a small portion of a sentence. Ma et al. proposed an emotion estimation system for chat or other dialogue domains based on keyword spotting with sentence-level processing [76]. In a more recent approach, Dey et al. proposed a rule-based model for emotion extraction in a real-time chat messenger based on a lexicon [77]. Another powerful example of rule-based systems is the Affect Analysis Model, which analyzes affect specifically in informal online communication media using symbolic cue analysis, syntactic structure analysis, word-level, phrase-level, and sentence-level analysis [78].

Lexical affinity approach uses the mutual information of words based on their relationships in the document. The aim is to link words that are relevant for certain affective dimensions and assign a probabilistic affinity value. Similar to keyword spotting, the disadvantage of lexical affinity is its inability to take the sentence level analysis into account, which makes it very limited in understanding complex and compound sentences.

Statistical NLP and learning-based approaches are popular, and they basically rely on automatic calculation of frequencies of some seed words, their co-occurrences, punctuations, abbreviations, and sometimes synonym and acronym information. Brooks et al. presented an automated affect classification system for chat logs exploiting NLP and machine learning techniques [79]. This system segments the chat data and uses an improved bag-of-words model (including non-verbal cues) to classify text into thirteen affect categories.

The commonsense-based approach was first proposed by Liu et al. for emotion classification [80]. They used three real-world commonsense databases, including David Lenat's famous Cyc [81]. Compared to other approaches, this model works robustly on the sentence level. In this approach, a number of emotion models corresponding to each emotion class compete with each other and the winning models are used to identify the affect label of the text segments.

The affect recognition literature from Turkish texts is quite sparse. Turkish is an agglutinative language, where words can take many suffixes that modify the meaning. Cakmak et al. analyzed emotions attributed to Turkish word roots and sentences, and found significant correlation between them based on valence, activation and dominance [82]. In their approach, they have used 197 Turkish emotion words. Boynuka-lın et al. analyzed emotion in Turkish texts by using machine learning methods [83]. Some recent NLP tools have been examined by Yıldırım et al. on a manually normalized set of 13K tweets in Turkish, for positive and negative sentiment analysis [84]. Vural et al. proposed a framework to classify the polarity of Turkish movie reviews. For this purpose, they translated the SentiStrength sentiment lexicon to Turkish and achieved 76% accuracy at word level and 75% accuracy at sentence level [85]. Dehkharghani et al. [86] presented SentiTurkNet, which is the first comprehensive Turkish polarity lexicon and included positivity, negativity, and objectivity scores assigned to each synset in the Turkish WordNet (about 15,000 synsets).

Our study was conducted on the most extensive social media text corpus in Turkish, and it targets continuous and dimensional affect prediction on valence, arousal, dominance (VAD) dimensions.

4.5.2. Data and Annotation Scheme

We develop our approach for the online chat domain, which typically involves limited vocabulary, grammatical irregularities, and chat-specific expressions, emoticons, and abbreviations. We use again COPA database of multiparty chat records, explained in detail in Section 3.

For manual annotation, we randomly selected about 1000 sentences from the 4 million sentences in the database. We selected the most expressive ones by excluding sentence fragments and meaningless utterances, obtaining 300 independent sentences (both emotive and non-emotive). Note that the final set is not necessarily a balanced distribution of emotive and non-emotive sentences. Then, we created nine pairs of surveys (each survey included 100 sentences) to evaluate for valence, arousal, and dominance dimensions. Sentence-level annotations were performed online and anonymously, by native Turkish speaker annotators from different backgrounds and ages. Each annotator was given a set of 100 sentences and asked to complete the survey for one dimension at a time after reading a single page of instructions and sample questions. For initial training, we also provided a set of tagged words drawn from the Affective Norms for English Words (ANEW) corpus [87].

Annotators were instructed to evaluate each sentence on a 5-point Likert scale, using their first intuition about the sentence. For valence; a value of 5 indicated extremely happy, satisfied, hopeful, pleasant, and 1 indicated completely unhappy, dissatisfied or bored. For arousal, the annotation ranged from calm, inactive, and dull at the low end of the scale to highly aroused, excited and active at the high end. Similarly, for dominance, the highest value was given if the subject felt powerful, dominant, influential or controlling, and the lowest value if they felt controlled, unimportant, weak, or influenced. For all dimensions, 3 was selected if they felt neutral.

In order to clarify the emotional dimensions and easily assess affective stimulus, we provided several annotated sentences as a reference and incorporated the Self-Assessment Manikins (SAM) [88] when designing the annotation scheme. Manikins were presented at the top of the sentences.

In order to perform a quantitative measurement for the effect of each modifier, we designed systematical annotation procedure for 72 intensifiers & diminishers, and 50 interjections that can shift the affective load[24] . This forms a second set of annotations, which we refer to as "modifier annotations". Rather than annotating modifiers at the

[24]All resources are available online. https://github.com/verdeerosso/affective-turkish

word level, we assessed the effect of each modifier in context, considering that they may have different impact depending on the emotional polarity of sentence. Therefore, we created two sets of 73 sentences. These two sets included the same wording and the only difference between corresponding sentences was the presence or absence of a modifier.

In all, we collected 6,300 ratings for sentence annotations and 7,575 ratings for modifier annotations. 116 participants completed annotation for one or more dimensions. The ground truth annotations for both sets were obtained by averaging the annotations for each sentence.

In order to achieve inter-annotator reliability, each sentence annotation was performed by seven different annotators with various educational and socioeconomic backgrounds in order to capture a broad consensus on affective judgment, which can be highly subjective. Then, we examined the inter-annotator agreement among all annotators. In addition to subjectivity, annotation is an error-prone process for several other reasons. Affective meaning can be ambiguous, or annotators might label the instances in a rapid manner, which possibly contributes to noise level.

For the annotation of modifiers, we repeated the labeling for 15 times on average with different participants. Repeated labeling is especially useful for noisy labels [89]. For both set of annotations, the values that were more than one standard deviation to the mean were treated as outliers and eliminated, and the means were recomputed.

There is no comprehensive and widely-used lexicon of affective words with VAD annotation for Turkish. Since it is very costly and time consuming to construct a dictionary from scratch, we have automatically translated a dictionary of English lemmas. The study of Warriner et al. [90] is based on the ANEW norms proposed by Bradley and Lang for 1,034 words [87]. They have rated 13,915 English lemmas in a nine point scale (1-9) and provided mean values and standard deviations for valence, arousal, and dominance scores. A total of 1,827 participants (through Mechanical Turk) contributed to their study.

We linearly transformed these affect scores to a five point scale [1-5]. For translation, we initially used the Google Translation API. Two independent human translators manually checked each word, and corrected missing words and mis-translations. Some culture and language dependent words are omitted and the affective lexicon is finally expanded with synsets from TDK (Turkish Language Organization) dictionary[25] . As a result, we obtained a comprehensive affective lexicon for Turkish that includes valence, arousal and dominance values of 15,222 words and phrases. In Table 4.1, we show some examples from our affective lexicon. While this resource is not entirely reliable and well-formed, our assessment shows that it is useful. Any future work on a proper Turkish affective lexicon would improve the system that we propose.

[25]http://www.tdk.gov.tr/

Turkish	**English**	**POS**	**Val.**	**Aro.**	**Dom.**
açıkhava	outdoor	ADJ	4.17	2.28	3.1
açığa kavuşturmak	clarify	VB	3.5	1.93	3.52
adaletsizlik	injustice	NN	1.73	3.73	2.14
adam kaçırma	kidnapping	NN	1.53	3.18	1.74
mutlu	happy	ADJ	4.74	3.53	4.11
tatil	vacation	NN	4.77	3.11	4.06
yetenekli	talented	ADJ	4.48	2.78	3.57
yumuşak başlı	docile	ADJ	3.38	1.74	3.23
polis	cop	NN	2.75	2.95	1.92
akordiyon	accordion	NN	3.13	1.97	3.11
bebek bezi	nappy	NN	2.05	2.16	2.62
bebek karyolası	cot	NN	3.19	1.98	3.04
donuk	dull	ADJ	2.2	1.34	2.86
yatıştırıcı	soothing	ADJ	4.03	1.46	3.38
zelzele	earthquake	NN	2.03	3.88	1.57
lenfoma	lymphoma	NN	1.8	2.8	1.69
ümit	hope	NN	4.24	3.15	3.89
ağlamak	cry	VB	2.11	3.23	1.78
yusufçuk	dragonfly	NN	3.73	2.43	3.21
centilmence	gentlemanly	ADV	3.89	2.46	4

Table 4.1. Some example words and phrases from our affective lexicon.

4.5.3. Model of Affect Analysis

We performed a three-step pre-processing method to normalize the informal characteristics of the chat messages. Before normalizing the noisy chat texts, the input sentences were segmented into words and each token was kept with part-of-speech (POS) information. We recorded the intentional spelling mistakes, duplications (e.g. 'selaaaam'), upper-case usage (e.g. 'HA-DII') or exclamation mark usage (e.g. '!!!'), since they serve as useful features, especially for high arousal patterns.

Secondly, we checked this list of corrected words with a direct look-up. Because there are a lot of repetitive expressions and repetitive words in chat domain, spelling mistakes are usually repeated too. Therefore, keeping a list of the most frequent normalizations is very effective to correct chat-specific spelling mistakes. Lastly, we used a Turkish normalization tool proposed by Torunoglu and Eryigit [91]. This tool is available online.[26]

People often use emoticons to enhance or underline the meaning of certain text elements, to enrich the quality of the conversation, and sometimes, just for fun. In order to have a comprehensive affect sensing, we examined textual messages for various affective attributes. Emoticons clearly reinforce the emotional communication in a powerful way and therefore we considered them to be most informative qualities in our model when they are present. In the simplest form of emoticon exchange, special character combinations are used within the context of online chat to display affective mood. Frequently used examples include ':o' as 'surprised', ':(' as 'sad', and ';)' as winking. We constructed an emoticon table consisting of 120 popular emoticons used in Turkish multi-party chats.

Modifiers and interjections make up an important set of features. We collected a list of Turkish words (adverbials, adjectivals, and nominals) that can intensify or diminish the affective attribute of a sentence. We enriched this set with the synonyms of these words. The final modifier list includes 72 intensifiers and diminishers. We also created a list of 50 interjections.

Since Turkish is a morphologically rich agglutinative language, one can generate hundreds of legitimate words from a single root with derivational and inflectional morphemes. Subsequently, even word-level analysis is challenging [92]. On the other hand, there is a strong link between word roots and the perceived emotion of the sentence [82]. Therefore, we take the **word roots** into consideration when we create our lexicon and for analyzing the affect in sentences. Assuming that affective load of a sentence is almost the same for different person forms and tenses, we do not perform

[26]http://tools.nlp.itu.edu.tr/Normalization

a detailed morphological analysis for these. Morphological operations that we do take into account include vowel harmony, such as consonant changes (e.g. ayak - ayağı), vowel drops (e.g. burun-burnum) [93], removing the infinitive suffix '-mek', '-mak', and most importantly, detection of negation that comes with the negation suffix '-me', '-ma'. We also checked other negation markers of Turkish, such as 'değil' (not), 'yok' (there is not), 'ne...ne' (neither .. nor), '-siz', '-sız', '-suz', 'süz' (without).

These features are prioritized in our system, which first checks for the existence of any emoticons, then looks for modifiers, surface features, negations, and finally, the corresponding VAD scores for each word unit.

Based on these features overall affective value is extrapolated as follows:

- If there is any modifier connected to a verb or a noun as phrasal, the score of the word is updated based on the polarity of the sentence and on the particular coefficient of the modifier. These coefficients are obtained from the modifier annotations.
- If there is a verb and a noun phrase with opposite scores, the verb is considered as dominant.
- If negation is detected, the VAD score is reversed by subtracting from 6 (e.g. 2.3 turns into 3.7). However, the 'ne ...ne' connector neutralizes the affective value of the sentence to 3.0 (e.g. "Sabah hava ne iyiydi ne de kötü." - "This morning the weather was neither good, nor bad.")
- If there is an emoticon and a word with a conflicting score (mostly the case for sarcastic and ironic sentences), the emoticon is taken as a reference.
- If there are no affect-carrying words, emoticons, or interjections, the sentence is considered as neutral.

The quantitative effects of a modifier on the valence, arousal and dominance were calculated by comparing the mean annotation scores with the presence or absence of each modifier. Each modifier was tested both in a positive context and a negative context, as we initially assumed (and then showed) that some modifiers might intensify

the valence when used in a sentence with positive polarity and diminish the valence in sentences with negative polarity. Therefore, when there exists a modifier, our model first determines the polarity (positive or negative) of a sentence, and then updates the affective scores based on the polarity assignment. We followed the same process for annotations of interjections and some other surface features, such as the effect of use of upper case typing, or intended spelling mistakes (e.g. 'helloooo') to boost the arousal score.

4.5.4. Evaluation of the Model

Based on our annotated affective data, we report fine-grained (dimensional) and coarse-grained (sentiment level) evaluation results with different metrics. For dimensional evaluation, model scores were scaled continuously between [1-5]. The results were reported for all three dimensions in Table 4.2, in terms of mean squared error and accuracy. To compute fine-grained accuracy, we calculated difference between the ground truth score (annotation score) and the predicted model score. If the difference was smaller than 0.5, the prediction was considered to be correct. To compute accuracy, the ground truth score and the predicted score are both rounded to the nearest integer, and a match is if they agree exactly. Since the scale has five intervals, random agreement has %20 accuracy.

Measure	Valence	Arousal	Dominance
All features (MSE)	0.62	0.58	0.47
All features (Acc)	%50	%54.2	%57

Table 4.2. The accuracy of the model for dimensional affect estimation.

As can be seen from Table 4.2, the mean squared error is lower for dominance and arousal compared to valence. However, the annotation score ranges are different for these three dimensions: Valence: 2.85($\pm$0.86); Arousal 3.06($\pm$0.92); Dominance 3.12($\pm$0.69). While the accuracy is lowest for valence, correlation tells a different story. The correlation between predicted scores and the ground truth annotation is highest for valence (0.62), lowest for arousal (0.25). Lower spread for scores makes

prediction more difficult, as it implies that the distinctions between sentences are not easy to discern, even for the human annotators.

Measure	**Accuracy (%)**
All features	70.4
All features except modifiers	57.2
All features except emoticons	68.4
All features except normalization	64.9
All features with SentiTurkNet	62.8

Table 4.3. The accuracy of the model for coarse-grained affect estimation.

Secondly, for coarse-grained evaluation, model predictions in valence dimension are mapped to positive (¿3) and negative (¡ 3) classes. This is the sentiment analysis scenario with results displayed in Table 4.3. We obtain %70.4 accuracy with this approach when all features are employed. In order to evaluate performance of our affective lexicon comparatively, we also tested the model with SentiTurkNet polarity lexicon [86] and obtained %62.8 accuracy with this setup.

4.5.5. Conclusions and Future Directions

In this section, we described a rule-based system for dimensional affect analysis in Turkish multi-party chat environment. We also presented a comprehensive Affective dictionary in Turkish language with Valence, Arousal, and Dominance scores in the [1-5] scale. The number of tools and prior work in analyzing affect from text is very limited for the Turkish language, and such a corpus will be beneficial. Partly due to the lack of such resources, there are no comparative results reported in our study. The affective lexicon and the other resources created in this study are made publicly available.

The application domain of the study is multi-party online social games. As indicated in [94], person-dependent models are more successful in affect recognition. A possible extension of the study is to learn person-dependent models by further employ-

ing contextual information (e.g. game data for multi-party chat during games). We have not used any machine learning approaches, but directly modeled frequency distributions of words and modifiers. For supervised machine learning approaches, more extensive annotation of sentences is necessary. The existing work can help annotation efforts by focusing them on parts of the data space where predictions are poor.

4.6. Authorship Recognition in a Multiparty Chat

In games, some users who are blocked by administrators for various reasons (such as cheating, foul language, hate speech, abusive behaviors) may return to the game using an impostor account. Finding these matching accounts is a very hard problem to tackle manually. Game communities spend resources to preserve a user friendly gaming environment, which includes containing offending players. Reducing the number of suspects might be very useful, even if finding the real offender is difficult.

Writing style is unique for everybody, and some identity-related cues remain even if the individual consciously attempts to change the writing style [95]. This issue was investigated in the context of authorship recognition, which seeks to identify the author of a piece of text from among a set of candidate authors, whose texts are available for supervised classifier training. The electronic chat domain is significantly different from the literary text domain. These differences are particularly prominent in word and character frequencies, use of punctuation marks, intentional and unintentional misspellings, vocabulary usage, sentence length, and the particular ordering of words. The increased freedom in the usage of language, coupled with (typically) much more limited vocabulary makes chat biometrics an interesting challenge.

We investigate the rate of success in identifying these malicious users in multi-participant chat environments by means of extracting relevant features and supervised classification techniques. In our approach, we apply and compare several methods to match users to a gallery by their chat records. In the literature, methods developed for matching personal text content have been mostly evaluated with Indo-European languages. We test our approach with documents that have Turkish chat content, which

bring additional challenges due to the agglutinative nature of the language (i.e. many postfixes are applied on word roots). Text analysis in agglutinative languages includes a normalization step to isolate the roots of the words, which we additionally assess in the context of chat biometrics.

The rest of the section is organized as follows. Section 4.6.1 overviews the problem and related work. Section 4.6.2 describes the proposed method for identifying a person given his or her chat records and baselines. Section 4.6.4 reports our experimental results on the COPA Database, as well as the results of the approach on a standard authorship identification benchmark as a sanity check.

4.6.1. Related Work

Every authorship recognition (or identification) problem contains a training corpus in which there is a set of text samples for candidate authors and a test corpus of text samples from unknown authors. Each sample should be attributed to a candidate author. Identification approaches can be distinguished as profile-based and instance-based, according to whether the set of text samples for each author is treated individually or cumulatively [96].

Concatenating training texts per author in one single text file is known as the profile-based approach. This large single file is used to extract properties of the author's style. A text sample from an unknown author is compared with each author profile, and a suitable distance measure is used to find the most likely author. In this approach, features related with the variety of texts in the training corpus are not taken into consideration.

Instance-based approaches, on the other hand, consider each text sample independently, hence the differences in the training texts by the same author are not neglected. Both approaches have their own advantages, but if text documents are very concise, concatenation of the text (as in profile-based approaches) may help to create a sufficiently long document for capturing the author's style [97]. Performance in this domain

depends on identification methods, as well as pre-processing techniques, document set sizes, language characteristics, and feature sets. In terms of used features, character N-grams, word tokens, term frequency - inverse term frequency (TF-ITF), distribution based similarity features are typically used. Features are categorized based on different attributes of text: lexical, syntactic, semantic and character based features, respectively [98]. Some of the most commonly used features are listed in Table 4.4.

There are a few important studies related to chat biometrics on texts in English. Inches et al. [99] used two different internet relay chat (IRC) datasets containing homogeneous and heterogeneous topics separately. Traditional chi-squared distance and Kullback-Leibler divergence were used to determine the similarity between the author profiles. The study achieved up to 61% accuracy on heterogeneous chat records. Layton et al. [100] used IRC records of 50 users, each of whom entered 50 chat messages. The re-centered local profile (RLP) method was used for identification. Using an ensemble classification scheme where each classification was weighted by the ratio between the distances to the second closest and closest authors, an accuracy of up to 55% was achieved.

If we consider simplicity and language independence as primary factors, character based features are expected to perform better. Especially, the character n-gram representation has been used as one of the most effective measures of authorship attribution [101, 102]. On the other hand, lexical, syntactic and semantic features have some advantages over each other. For example, superiority of syntactic and semantic features depends on the idea that authors tend to unconsciously use similar patterns, and some language-specific NLP tools (such as a POS tagger, stemmer, spell checker) may be required for extracting syntactic and semantic patterns.

The bulk of authorship analysis approaches in the literature focus on English language, which is weakly inflected. Despite the fact that research on highly inflected languages like Greek and Sanskrit or fusional languages like German may be useful in order to understand language dependent approaches on chat biometrics to some extent [97, 103, 104], agglutinative languages differ from them by a complex word

structure, which is formed by stringing together morphemes without changing them in spelling or phonetics. An example of such a language is Turkish, where for example, the word *evlerinizden*, or "from your houses", consists of the morphemes, *ev-ler-iniz-den* with the translation of *house-plural-your-from*.

There are some prior studies on author attribution in Turkish. Tufan et al. used style markers as features, on a gallery of 20 authors [105]. Amasyali et al. used n-gram model in text categorization for author, genre and gender classification [106]. Both studies used corpora collected from newspaper articles, which are written in formal Turkish.In another study, a chat mining framework was tested on a Turkish dataset containing peer-to-peer text messages [107]. This work is one of the most exhaustive efforts on chat biometrics in Turkish, and while it does not cover multiparty chat, it established that context plays a significant role in style. However, term-based features achieved better results compared to style-based features on a 100-author problem.

4.6.2. Methodology

For this study, we utilized COPA dataset chat logs. The COPA database is particular in that messages are always written in a multiparticipant fashion (there are always four players in a game); they are unedited (except for a black-list that contains the most frequently attempted insults); and they are spontaneously produced. The number of chat and game records per player vary greatly. Consequently, we have pre-selected a subset of the dataset for the problem of chat biometrics before any research or modeling took place. We sorted chat participants according to the number of unique words used by each, and eliminated participants who had vocabulary sizes less then 100 unique words. This is a very coarse pre-processing, but people with very limited vocabulary might be easier to identify, and might positively bias the results. The remaining users are sorted in decreasing order according to number of active chat sessions, and the most active users are selected for building a chat biometrics benchmark database. With 978 users, this database is one order of magnitude bigger than the most relevant work from the literature. Table 4.5 describes the properties of the final database. Since we use 5-fold cross validation, as well as to assess the effect of

the number of chat entries per user, we required that at least 5 chat sessions per user should exist for each of five folds.

We adopt the re-centered local profile (RLP) approach, proposed by Layton et al. [108], which uses a language profile in the calculation of a distance between an author and a document:

$$d(f_1, f_2) = \sum_{n \in profile} [f_1(n) - P(n)].[f_2(n) - P(n)]$$

where f_1 and f_2 are author/document profiles to be compared and P is the language profile, which is extracted from the entire training set as an approximation to the absolute language profile. If we normalize profiles by using absolute distance of variation between each profile, the following equation is obtained:

$$d(f_1, f_2) = \sum_{n \in profile} \frac{[f_1(n) - P(n)].[f_2(n) - P(n)]}{|| \ f_1(n) - P(n) \ || \ . \ || \ f_2(n) - P(n) \ ||}$$

Because of the flexibility inherent in natural languages, extracting the absolute profile of a language is impossible. For this reason, all the normalized author profiles in the training set are combined to extract a standardized language profile.

The common n-grams (CNG) method proposed by Kešelj et al. [103] uses the relative distance between two documents (or author profiles), and serves as a basis for RLP. However, the most noticeable difference is that RLP measures the profile similarity according to most distinctive features, rather than the most frequently used features, using the standardized language profile approach described above.

For each entry, we have replaced capital letters with small cases. Then, $N \times N$ sparse bi-gram matrices for each user are calculated. We tested N equal to 32 and 66 (See Table 4.6) based on most common characters used in the COPA database. In addition to the RLP approach, we tested Cosine Similarity (CS) based identification.

4.6.3. Experimental Protocol

In our dataset, 978 active users (from among more than 30,000 users) were selected in order to set up an experimental protocol. Accumulated chat records of a user in one session were used as an instance for that user in the corpus. 5-fold cross validation was used.

To assess the number of characters and the number of words needed to identify a person, we performed two different sets of experiments. The entire set of users were labeled as SET 1. A minimum number of 140 characters per session was required in SET 2, which subsequently has only 617 users.

During testing, we have systematically increased the number of instance per test user from 1 to 100 in order to understand how performance changes with respect to the number of samples per user.

We describe the experiments as abbreviations in the tables, formed of identification method (CS or RLP), extracted n-gram features (32 or 66), and whether a minimum number of characters was required per session or not (GT140 for required and GT0 for not). For instance, RLP/66/GT140 means that re-centered local profile is used as identification method, features are extracted as 2-gram character matrix of size 66×66, and chat sessions in SET 2 were used, with at least 140 characters per session.

A small part of the raw data on COPA was normalized by using the web API of the Turkish NLP tool [109], whereby intentional or accidental misspellings were replaced with correct forms. Since Turkish has flexible sentence structure, as well as agglutinative word forms, the normalization affects the identification performance significantly. For instance the raw sentence "*büttttttüüüünnnn insnlar e$it dogaaarr*" ("all people are born equal") is correctly normalized as "*bütün insanlar eşit doğar*". Normalization changes the distribution of the n-gram features. In the next section, we report results with both raw and normalized test data. For the latter, the normalized

version of training data was used.

4.6.4. Results

In the first experiment, RLP and CS measures are used with the 2-gram character matrix size of 66×66 for each user. Given sufficient data for processing each user (results on SET 2), RLP and CS do not give significantly different results, as shown on Table 4.7. We tested for significance of differences with paired t-tests between the identification results. Conversely, if all chat sessions are taken into consideration, a significant difference was observed between CS and RLP, with RLP being the more accurate approach. Increasing the gallery size has a more detrimental effect on CS compared to RLP, but we should remember that the added users (i.e. those that are present in SET 1 but not in SET 2) are the ones with shorter utterances.

A second experiment was conducted to inspect the effect of the number of chat returns (entries) of a user per test case. *SET 1* was used, and *RLP/66/GT0* is the protocol for this experiment. The number of entries per user was systematically varied between 1 and 100, and an accuracy of 77.5% was obtained with 100 entries. The curve stays flat after the 92nd entry, suggesting that adding much more test data may not result in obtaining better results. Figure 4.3 illustrates the relationship between the number of chat entries and the rank-1 identification rate.

One of the issues we investigate with this study is how text normalization impacts author identification from Turkish chat records. Our results show that raw chat data is more distinctive than normalized chat data, since intentional misspellings or unconscious typos are some of the most important features for identification. Normalization of text causes loss of these distinguishing features. The impact on the results is evident in Table 4.8, which reports the raw and normalized versions of each test setting. By performing a paired t-test, we also confirmed that the difference is statistically significant with $p < 0.0001$. We used a small set of users for the normalization experiments, and since the effect was very clear, we did not perform normalization on the entire set of users.

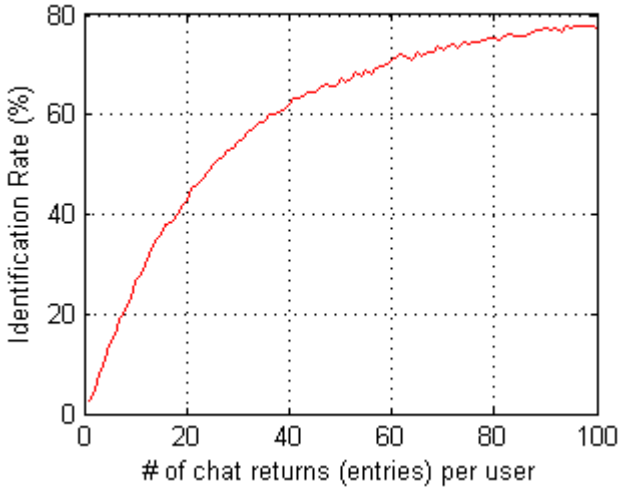


Figure 4.3. Change of identification rate for test type *RLP/66/GT0* on *SET 1* while increasing the number of chat entries per user from 1 to 100 during testing.

We have contrasted several approaches proposed for authorship identification on the problem of chat biometrics, and performed tests with a large database of multiparty chat records in Turkish. While normalization is a standard step in text processing for agglutinative languages, we illustrated that it results in accuracy loss. Our results show that it is possible to obtain around 75% rank-1 accuracy for a gallery size of 978.

4.7. Machine Learning Overview

Several classifiers were studied as possible candidates to be employed for the machine learning part of our work. In this section, we will briefly overview each method that we evaluated. First, three of the common machine learning algorithms were implemented and evaluated; Kernel Support Vector Machines, Decision Trees and Naive Bayes. Next, we focused further on two machine learning methods that gained popularity in recent years, Bayesian Point Machine (BPM) and Gradient Boosting Machine (GBM). We will be presenting the performance of these methods in later sections with experimental results.

4.7.1. Kernel Support Vector Machines (Kernel SVMs)

Support Vector Machines (SVMs) are a set of supervised learning methods which can be used for both regression and classification. A linear SVM model is composed of sets of support vectors and weights which manage to divide the sample space by a clear gap. After supervised training, new examples are mapped into the same space and classified based on the part of the space they fall in to. Output of a given SVM with N support vectors of z_i and weights w_i is calculated by the following formula;

$$F(x) = \sum_{i=1}^{N} w_i \langle z_i, x \rangle + b \tag{4.7}$$

Note that SVMs are linear classifiers in origin [110], which means they can only distinguish linearly separable clusters of examples in feature space. However, using the kernel trick first proposed by Aizerman et al. [111], non-linear classifiers can also be formed using SVMs. The resulting algorithm is essentially similar to the original SVM, except that dot products are replaced by a non-linear kernel function. This allows to form a transformed feature space which can manage the classification better by forming a non-linear but clearer gap in between classes. The transformation may be non-linear and the new space may be high dimensional, but the hyperplane in between classes are still linear in nature. However, the linear hyperplanes in the transformed space may represent a non-linear boundary in the original input space.

There are some common kernel functions such as linear, polynomial, Gaussian and Sigmoid. In this study, we are evaluating the three of these kernel functions and report their performances with our dataset. For learning, we use sequential minimal optimization [112].

4.7.2. Decision Trees

Decision trees use simple conditional rules repeatedly to map input feature vectors to a target output value. As a side benefit, the resulting decision tree after training provides humans an understandable structure about the solution. The tree structure is simple to traverse with the data which may provide insight and an explanation about the solution.

There are two popular learning algorithms for training decision trees; ID3 [113] and the C4.5 [114], both of which are greedy algorithms that result in nodes with local optimum decisions to arrive at a generalized tree topology. Both algorithms favor simpler and smaller trees for more generalization power. Note that, in our experiments, for training the decision tree, we utilise C4.5 algorithm which can handle continuous variables as features as opposed to ID3.

In decision tree algorithm, the minimization of error is handled using an ordering criteria with information gain. Using information gain, the algorithm decides which feature to use next by calculating the total loss of information when a feature is left out. Unfortunately, the algorithm is known to have a bias for features with high number of distinct values [115]. In order to overcome this, one can select gain ratio as an alternative criteria [113], which is defined as:

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)} \tag{4.8}$$

$$SplitInformation(S, A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \left(\frac{|S_i|}{|S|} \right) \tag{4.9}$$

$SplitInformation$ term in Formula 4.9 is used to minimize the importance given to features with uniformly distributed values.

Unfortunately, both learning algorithms rely heavily on heuristics and they cannot guarantee to produce smaller and more general structures. Furthermore, the problem of finding the optimal solution is known to be NP-complete [116]. On the positive side, decision trees perform extremely fast to evaluate. In addition, for problems that have a feature vector where binary decisions are possible for each feature, decision trees can achieve satisfactory results. Our classification problem meets these criteria, which made decision trees a good candidate to evaluate in our studies.

4.7.3. Naive Bayes

Naive Bayes (NB) classifier is a probabilistic classifier with an independence assumption where every feature in the feature vector is considered to be independent from the others. Using supervised learning, NB classifiers can be trained very efficiently, generally using maximum likelihood for parameter estimation in their Bayes models.

Although NB classifiers simplify the problem with the independence assumption, they have proven to be a well performing tool in complex machine learning problems [117]. In addition, they require a small amount of training data to estimate the internal parameters. They can also be used with imbalanced datasets [118, 119]. However, a comparison with more recent approaches such as boosted trees or random forests show that NB is not the best performing algorithm [120].

4.7.4. Bayes Point Machine (BPM)

BPM, first proposed by [121, 122], are non-linear kernel classifiers used for binary classification. We briefly describe this scheme in this section, and refer the reader to the original paper for a more extensive treatment of the mathematical foundations of the approach.

The BPM essentially approximate Bayesian inference for linear classifiers in the kernel space. A kernel is a transformation of training samples to a higher dimensional space. Similar to SVMs, once data is projected to a suitable space, a hyperplane that separates the data linearly can be estimated. The classification is achieved by determining the loss incurred by a class assignment, weighting it according to the posterior probability, and selecting the class label that achieves the minimum expected loss.

BPM approach improves the accuracy of the center of mass of the version space, the set of possible solutions that are consistent with the observed training examples, when compared to other kernel classifiers such as SVMs [123]. [122] also show that SVMs can be derived as an approximation of Bayes-point classifier. A more detailed comparison of BPM with SVM can be found in [124].

4.7.5. Gradient Boosting Machine (GBM)

GBM, first proposed by Friedman [125], is a popular machine learning technique that is based on ensembles of sequential weak learners. GBM minimizes an error term based on classification error using a gradient descent method iteratively.

More formally, for an estimation problem with input feature vector $\mathbf{x} = \{x_1, \ldots, x_n\}$ and an output variable y, a supervised learning method uses a training set of N samples of known $\{y_i, \mathbf{x}_i\}_1^N$ values. The algorithm seeks a function $F^*(\mathbf{x})$ which maps an input $\mathbf{x}$ vector to output y, such that for a distribution of $(y, \mathbf{x})$ pairs, the expected value of a loss function $\Psi(y, F(\mathbf{x}))$ is minimized:

$$F^*(\mathbf{x}) = \underset{F(\mathbf{x})}{\operatorname{argmin}} E_{y,\mathbf{x}} \Psi(y, F(\mathbf{x})) \tag{4.10}$$

GBM approximates $F^*(\mathbf{x})$ with a set of additive ensemble of weak learners of the

form $h(\mathbf{x}; \mathbf{a})$ which are simple functions of $\mathbf{x}$ with parameters $\mathbf{a} = \{a_1, \ldots, a_n\}$:

$$F(\mathbf{x}) = \sum_{m=0}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{4.11}$$

The expansion coefficients β_m are estimated with training data along with the parameters $\mathbf{a}_m$ in an iterative manner. The algorithm is initialized with randomly assigned coefficients, and the initial estimate is refined in M steps:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{4.12}$$

Only a subset of the training samples are used in each intermediate training step to inject randomness and prevent overfitting. Each weak learner in step m is trained by using to the total pseudo residuals from the previous step:

$$r_{im} = - \left[\frac{\partial \Psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F(\mathbf{x}) = F_{m-1}(\mathbf{x})} \tag{4.13}$$

GBM is highly configurable, with several options for weak learners h. For a recent overview of GBMs, see [126]. In our approach, we use decision trees, as they can be easily interpreted. For tuning the internal parameters without overlearning the training set, a 5-fold cross validation scheme was employed.

One of the powerful features of GBM is the possibility of inspection of actual contribution of individual features upon training. The algorithm allows evaluating each

feature to the overall prediction accuracy. In our case for evaluating game related complaints, such information may help human moderators when they start investigating these incidents manually.

4.7.6. Implementation

In our study, we have mainly used R programming language and *caret machine learning R library* [127]. The library includes several machine learning methods including SVMs, Naive Bayes, Decision Trees and GBM with extended packages[27] . The source code developed for our studies is made available as an Open Source GitHub public repository[28] .

For our experiments with BPM, we use the C# programming language with Infer.NET library [128] which incorporates expectation propagation algorithm [129] for the training of BPM. Note that Infer.NET is developed by the team that also proposed BPM algorithm.

4.7.7. Conclusion

There are several supervised machine learning methods in the literature that can be used for classification purposes in this setting [130]. An ideal approach for our problem should produce interpretable information for the human moderators, as the decision will invariably rest with them. Subsequently, black-box approaches are not preferable. In this study, we use Bayesian Point Machines (BPM) and Gradient Boosting Machine (GBM) for classification as our primary classifiers. We gave a brief overview of these methods, and explain our motivation for their usage in this specific problem. We have contrasted BPM and GBM with several other classifiers that produce interpretable results, including support vector machines, naive Bayes, Decision Forest classifiers. We report our performance and comparisons of these methods when applied to our dataset in Section 5.

[27] `http://cran.r-project.org/web/packages/gbm`
[28] `http://github.com/koraybalci/complaint-classification`

Lexical Features	**Character Features**
-total # of words	-total # of characters
-total # of unique words	-ratio of alphabetic chars.
-ratio of short words	-ratio of upper case letters
-mean word length	-ratio of digit characters
-mean sentence length	-ratio of white space chars.
-mean paragraph length	-ratio of punctuation chars.
-ratio of distinct words	-ratio of distinct chars.
-# of hapax legomena	-ratio of emoticons
-# of hapax dislegomena	-ratio of char. repetition
-word n-grams	-character n-grams
-skip-grams	-vowel combination
-word frequencies	-vowel permutation
-# of words of each length	-compression methods
Syntactic Features	**Semantic Features**
-freq. of function words	
-freq. of punctuation marks	
-part of speech (POS) tags	-synonyms of words
-total # of lines	-hypernyms of words
-total # of sentences	-semantic dep. graphs
-total # of paragraphs	-latent semantic analysis
-# of sentences per paragraph	-systemic func. grammar
-# of words per paragraph	
-# of characters per paragraph	
-ratio of spelling errors	

Table 4.4. Commonly used features for authorship recognition.

Corpus Characteristics	**Value**
# of users	978
# of chat sessions per user	261
# of chat returns per user	3,251
# of unique words per user	2,375
# of words per user	10,933
# of letters per user	39,494
# of capital letters per user	149
# of emoticons per user	288
# of digits per user	162
# of punctuations per user	679

Table 4.5. Statistics of the chat biometrics subset of the COPA database.

	Characters
2-gram of 32 char	abcçdefgğhıijklmnoöprsştuüqwxyz
2-gram of 66 char	abcçdefgğhıijklmnoöprsştuüqwxyz 1234567890 ”!’^+%&/()=?_*-<>—@:.,;‘

Table 4.6. Selected characters for 2-gram feature matrix

Database	**Test Type**	**Scores (%) with # of Chat Sessions**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SET 1	RLP/66/GT0	17.73	29.41	41.84	50.92	57.46	62.84	66.93	71.29	73.07	75.66
	CS/66/GT0	16.20	27.42	39.35	48.90	54.97	59.75	64.72	67.93	71.31	74.40
SET 2	RLP/66/GT140	32.22	54.13	66.22	74.39	78.61					
	CS/66/GT140	31.53	54.32	66.00	74.72	79.06					

Table 4.7. Comparison of character threshold effect on models. The RLP method is significantly better than CS for the first set (t-test, $p < 0.0001$), however, the second set does not show a significant difference. Short utterances contain relevant information, which gets lost in the second set.

Test Type	Test Data	Scores (%) with # of Chat Sessions									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CS/66/GT0	Raw	33.25	54.22	65.78	72.29	80.00	82.40	85.06	92.05	91.57	93.97
	Normalized	24.10	37.83	50.56	58.80	63.37	69.40	72.53	80.48	78.55	80.72
RLP/66/GT0	Raw	33.01	55.18	66.75	71.81	82.89	82.65	87.23	92.53	93.01	93.98
	Normalized	23.61	38.07	53.25	59.28	62.41	67.47	71.33	78.80	81.45	80.48

Table 4.8. Performance comparison of raw and normalized data of 83 chat users. The raw data produces significantly better results (t-test, $p < 0.0001$) compared to normalized data. This shows the textual errors are revealing in determining identity.

5. Experimental Results

5.1. Abusive Player Classification

5.1.1. Preliminary Study

In order to cluster players and analyze the complaints better, we try to incorporate all social signals used in the game to profile a player's characteristics. To achieve this, our feature vector for the player consists of the gender and number of several profile aspects, namely, the numbers of daily logins to the game, games played, game tables entered, wins, incomplete games, social rewards obtained, real life payments, virtual gifts sent, number of virtual friends, number of private messages sent, number of chat sentences and bad language attempts. We also use ratios of some of these properties listed above. For instance, in our feature vector we use the division of total chat sentences, wins, incomplete matches with player's total match count. As a result, we form a feature vector with 12 dimensions. We use 1.060 players for training BPM classifier and test with 100 cases.

In both training and testing groups, half of the players are complaint reporters and the other half are accused players. During the test phase, we systematically change the confidence threshold to observe the effect of the threshold on sensitivity, precision and specificity. This parameter will ideally be optimized on a validation set, and during the operation of the system, periodically updated to draw upon ever larger sets of training data.

5.1.2. Experimental Results

In Figures 5.1 and 5.2, the results for our tests with various confidence thresholds are presented. In Figure 5.1, the number of false and true positives for different thresholds are depicted. In Figure 5.2, the precision, sensitivity and specificity scores are given.

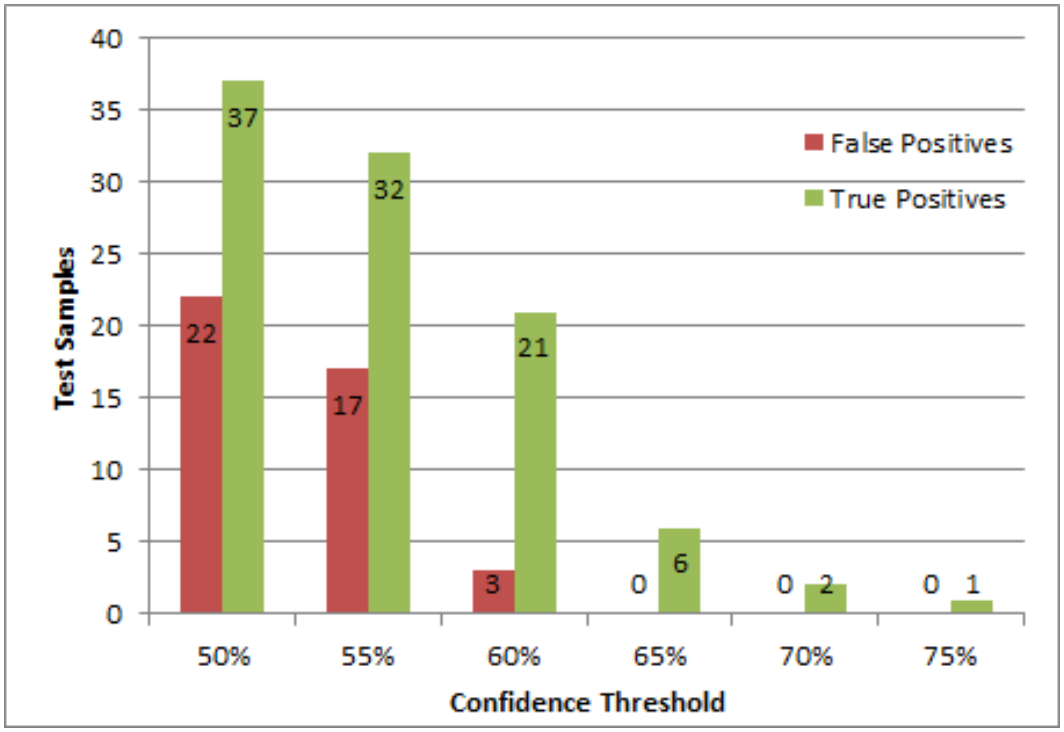


Figure 5.1. Players marked as offenders vs. different confidence thresholds.

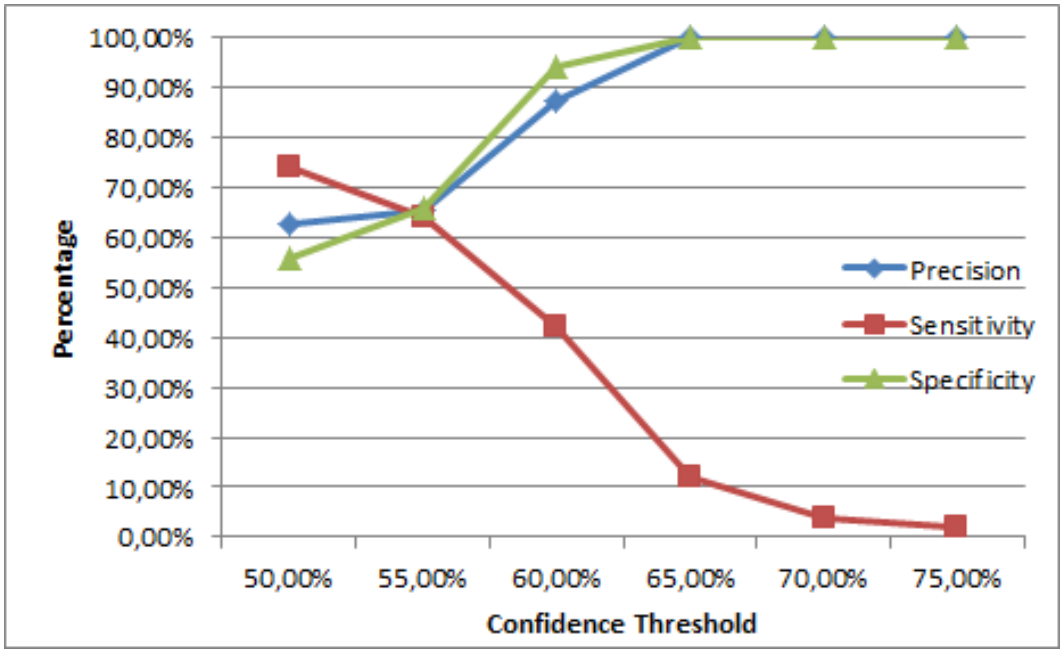


Figure 5.2. Precision, sensitivity, specificity vs. different confidence thresholds.

The results establish that differentiating a player accurately as an offender is a difficult task, and sensitivity values are low in general. However, if we increase the threshold for confidence, precision and sensitivity increases drastically. High precision means that if our method labels a player as offender, he or she is very likely to be a genuine offender. On the other side, high specificity says that if an innocent player has been reported as offender, we correctly conclude that he or she is innocent in most of the cases. A threshold value of 60% is sufficient for attaining 87.5% precision and 94% specificity, which are good hints for human moderators when evaluating a complaint.

In this preliminary study [17], we presented our analysis of player complaints data acquired from a real online social game. Our primary aim was to spot offenders by profiling players according to their in-game behaviour and performance. We propose a feature vector for player profile and a binary clustering methodology, followed by an adaptive thresholding for confidence to classify players reported in complaints as genuine offenders. Our approach performs well enough to aid human moderators in terms of precision and sensitivity so that they can prioritize and schedule which complaints to focus. However, by looking at the sensitivity scores, we observe that our methodology fails to classify all players in a fully automatized manner. Therefore, human moderators are still needed for precise judgments.

5.1.3. Exhaustive Study

First, we present the results obtained when all features are fed to BPMs individually and observe the effect of changing the threshold value, which is used to decide whether a player falls into abusive players category or not (See Fig. 5.3). Variance of precision, specificity and sensitivity for different threshold values show that increasing the threshold beyond 60% does not generate any significant gain for precision and specificity. However, sensitivity drops drastically as the threshold increases.

These results establish that differentiating a player accurately as an offender is a difficult task, and sensitivity values are low in general. A low sensitivity score states that our method can fail to identify all the abusive players. Meanwhile, high precision

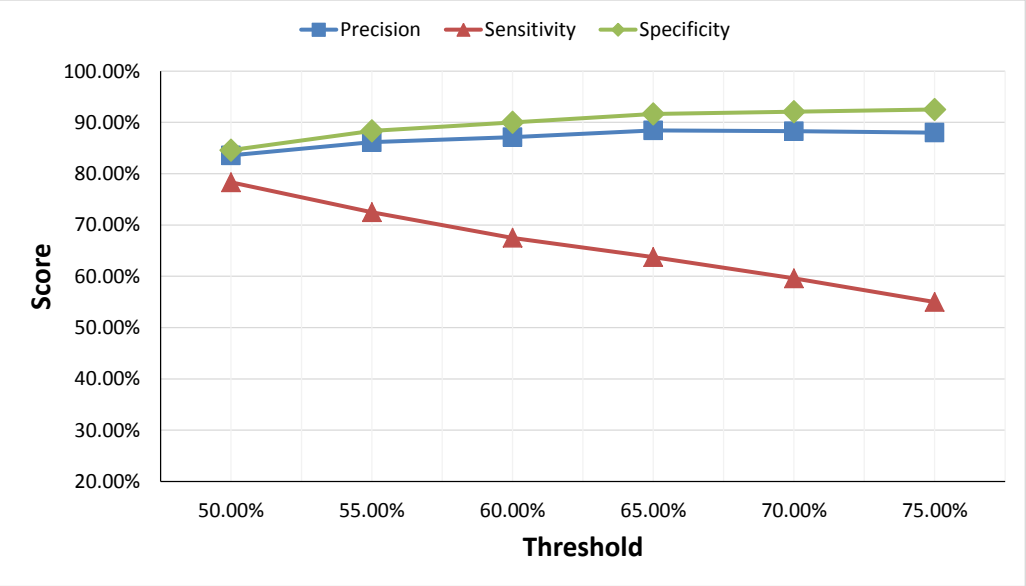


Figure 5.3. Results with all features fed to BPMs directly.

means that if our method labels a player as offender, he or she is very likely to be a genuine offender. On the other hand, high specificity says that if an innocent player has been reported as offender, we correctly conclude that he or she is innocent in most of the cases. As our primary aim with this study is to aid human moderators to evaluate submitted player complaints, choosing a high threshold value will assure less false positives and help moderators prioritize their time on complaints most likely to be caused by genuinely abusive acts.

To evaluate the contributions of individual features, we remove each feature from the analysis and inspect the deterioration in the system. Table 5.1 shows the loss of sensitivity in the absence of any given feature using 10-fold cross-validation results with a threshold value of 60%. Note that, we only report the features which produced a significant loss when they are left out from the *combined* feature set. According to these results, the factors with the greatest impact on sensitivity (i.e. for recognizing truly abusive cases) are gender and the daily login count. Our results show that male players have a greater probability than females for being an offender, which is not surprising. In Turkey, as in almost all other countries, the number of crimes and offenses perpetrated by males is vastly more than those perpetrated by females. The statistics published by the Ministry of Justice (2008) for numbers of offenders in law

suits in the categories related to verbal aggression illustrate this point: the cases for Threat involve 81.527 male vs. 13.039 female offenders, Obscenity involves 1.530 male vs. 85 female offenders, and Maltreatment involves 3.295 male vs. 471 female offenders. The daily login count is negatively correlated with aggressive behavior, which indicates that prolonged usage of the game reduces the probability of a user being aggressive.

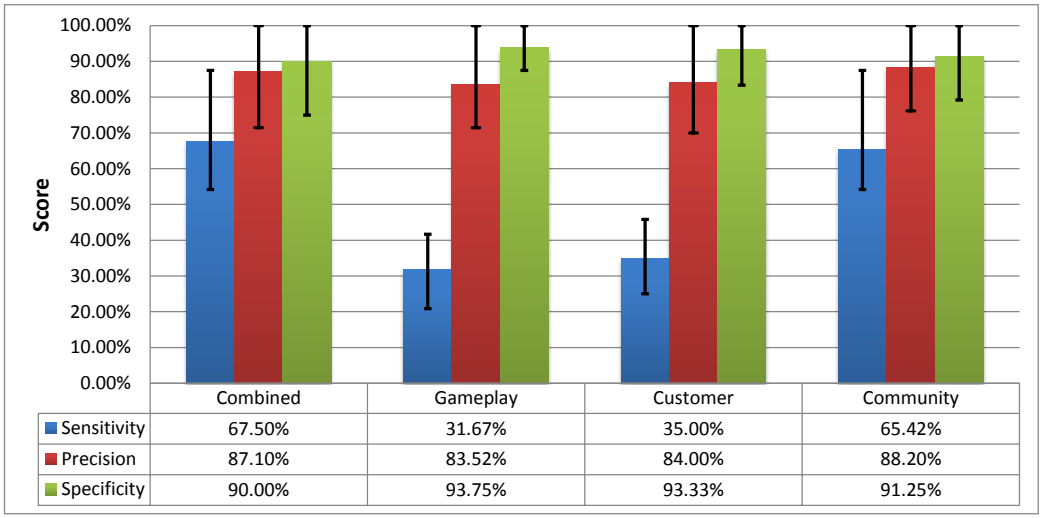
To assess the contribution of each feature to the overall sensitivity, we ran a series of experiments. In each case, we left out one feature and observed the drop in sensitivity with respect to the *combined* feature set. In Table 5.1, we report the most important features, and their contribution to the sensitivity measured thus. We also report significance results for the difference in the *combined* and reduced feature sets, obtained by one-tailed t-tests. Finally, we combine these informative features, and discard the rest, naming the new feature set as the *optimized* feature set. With the optimized feature set, we observe a sensitivity increase of 3.88% compared to the *combined* feature set, since misleading or uninformative features are removed.

Feature	**Contribution to Sensitivity**
Gender*	0.83%
Daily Logins*	0.83%
Bad Language Attempts*	4.58%
Friend Count	0.83%
Optimized Features	-3.88%

Table 5.1. The features that contribute to sensitivity, their contribution with respect to the entire feature set. * denotes significance at $p < 0.05$

Next, we evaluate our classifiers on different feature sets we have established previously. In Fig. 5.4, sensitivity, precision and specificity scores for each feature set are depicted. We chose 60% for the threshold. Note that the experiments for each set are repeated ten times using the 10-fold cross validation scheme, and maximum and

minimum scores of each experiment are plotted as error bars using thin lines overlaid on each bar.



	Combined	Gameplay	Customer	Community
Sensitivity	67.50%	31.67%	35.00%	65.42%
Precision	87.10%	83.52%	84.00%	88.20%
Specificity	90.00%	93.75%	93.33%	91.25%

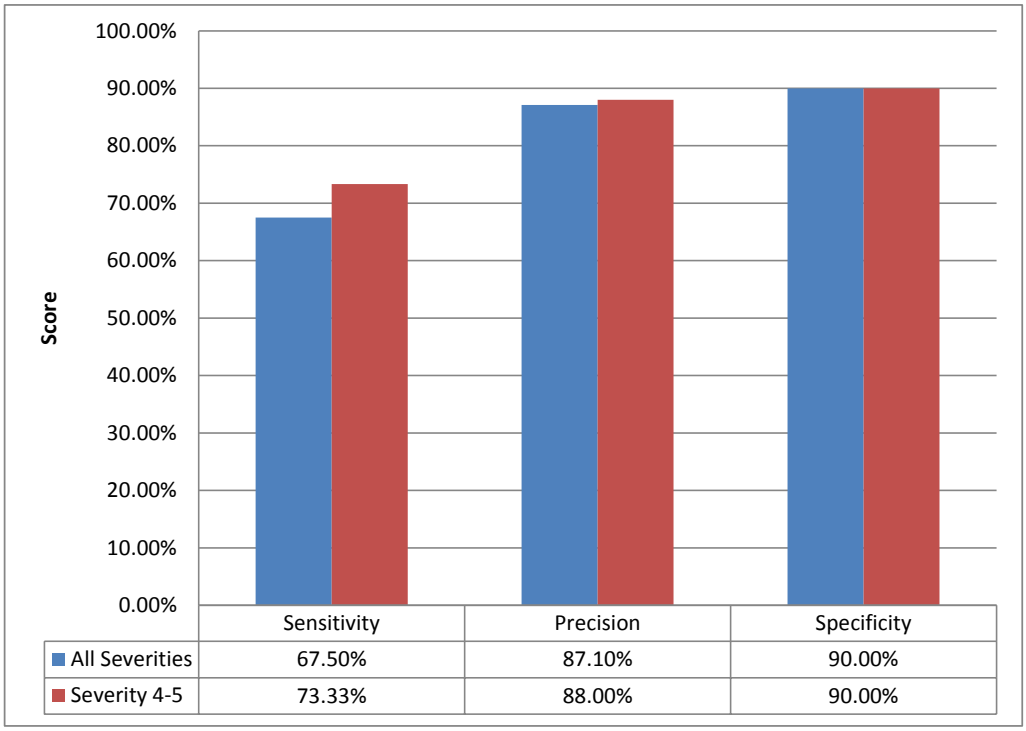
Figure 5.4. Scores for different feature sets. All sets have significance at $p < 0.05$

As shown in Figure 5.4, all feature sets perform well in terms of specificity (above 85%). This means that, for all classifiers, we create minimal false positives. The number of false positives for the *combined* set is the highest, but note that this set also achieves the highest sensitivity score (high true positives). The *optimized* set on the other hand, has the best precision, a sensitivity similar to the *combined* set, and a specificity that compares favorably with other sets. The resulting system confirms our initial claim that a well prepared system with a distinctive set of features can identify abusive players, and help human moderators effectively.

The *community* feature set performs remarkably well in terms of precision and sensitivity when it is used alone. This set also preserves sensitivity better, when compared with *gameplay* and *customer* feature sets. Therefore, we can conclude that social features play a very important role in detecting abusive players and just by looking at the social activities of a player can be sufficient to evaluate a complaint.

5.1.4. Abuse Severity

It is possible to adjust the system to ignore mildly abusive cases. Our annotations include a subjective judgment of the severity of each abusive case, denoted on a Likert scale (1 being the least severe, and 5 being the most severe). We have evaluated the performance of our approach when only complaints with high severity (4-5) are taken into account, both for training and testing. Fig. 5.5 summarizes the results. We obtain gain in both sensitivity and precision. This means that subjective evaluations of severity are congruent with the features we have proposed for the analysis, and the system indeed detects severe cases with higher reliability.



	Sensitivity	Precision	Specificity
All Severities	67.50%	87.10%	90.00%
Severity 4-5	73.33%	88.00%	90.00%

Figure 5.5. Score comparison for high severity cases vs all cases.

In Fig. 5.6, we report the change in sensitivity (ratio of true positives in the test set) for the complaints with severity of 4 and 5. Here, all abusive players (i.e. cases with all severity values) are used in training. Our results show that the system is capable of spotting severe cases with higher reliability, regardless of the threshold value.

We repeat our abusive player classification methodology for the new machine

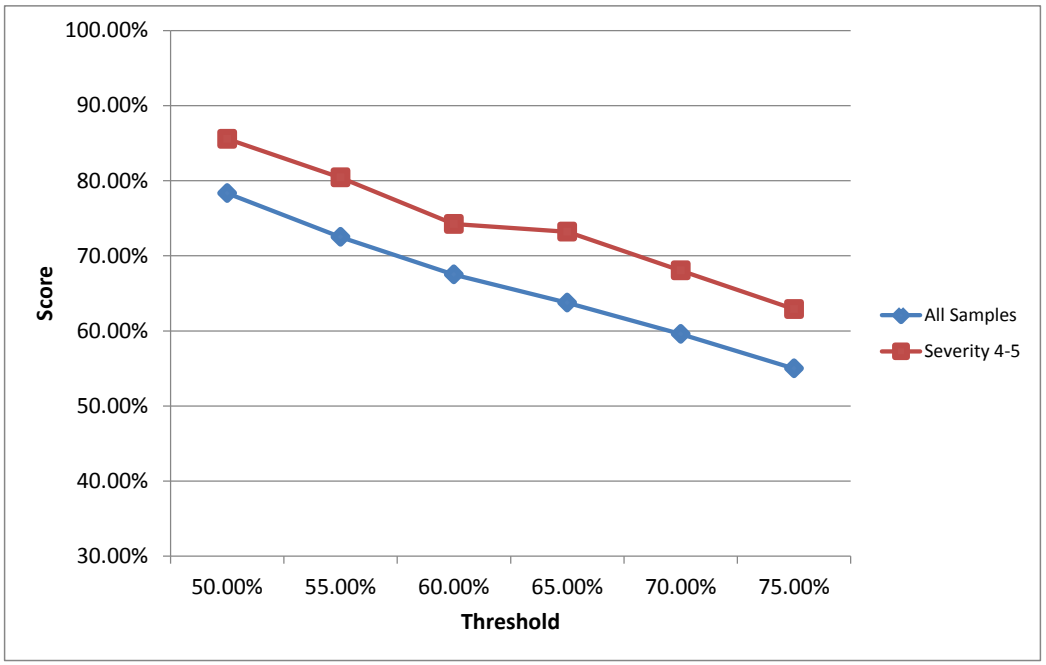


Figure 5.6. Scores for all samples vs samples with severity 4 and 5 under the *combined* feature set.

learning algorithms. The results are shown in Figure 5.7. Note that, we also include our prior results for BPM for comparison.

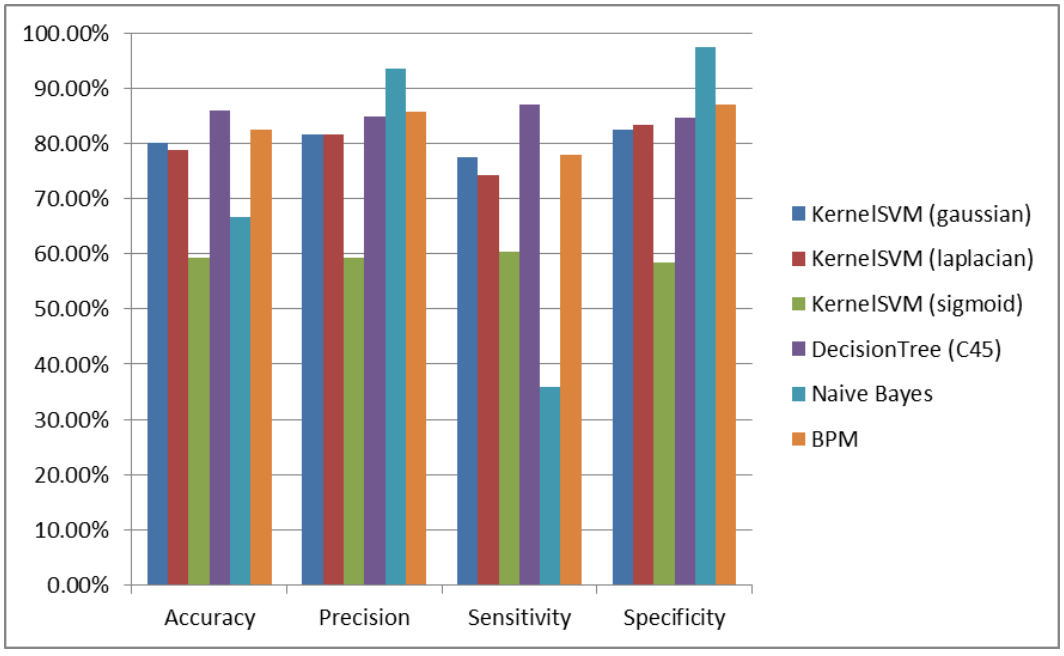


Figure 5.7. Results of abusive player classification with different classifiers.

To summarize, Decision Tree algorithm outperforms in terms of accuracy and sensitivity while Naive Bayes has better scores in terms of precision and sensitivity

while BPM performs comparably fine in all metrics.

5.2. Complaint Classification

Before proceeding with machine learning procedure, we statistically analyze the features to better understand the involvement of each feature.

First, we find and remove the features with near zero variance. In consequence, 'Credit Purchases' feature is removed from the feature set.

We studied highly correlated features and found 'Games Played', 'Wins' and 'Daily Logins' for both accuser and suspect being highly correlated with each other when a cutoff value of 75% is used.

Similarly, we also explored relevance of each feature to classification by applying recursive feature elimination (RFE) method with random forests using a 10-fold cross validation scheme [131]. The results of this experiment is displayed in Figure 5.8. The algorithm marked 27 contributing features as statistically relevant, with top 5 most contributing features as 'Gender', 'Bad Language Attempts', 'Games Played', 'Wins' and 'Chats' of suspected player. We keep the above features in the data, since machine learning algorithm of our choice for this study inherently evaluates each feature and removes the unnecessary ones automatically as well as providing means to calculate each feature's individual contribution to overall classification.

In Figures 5.9 and 5.10, we report accuracy, specificity and sensitivity for both configurations, using the training and holdout set respectively. Our results show that the proposed system reliably classifies player complaints containing abusive behavior. Including accuser profile and the information related to the recent communication of suspects and accusers increases the success rate. In this setting, the approach proposed in [18] achieves 62.61% accuracy with only suspect profile features.

The system detects severe cases with higher reliability, which will help human

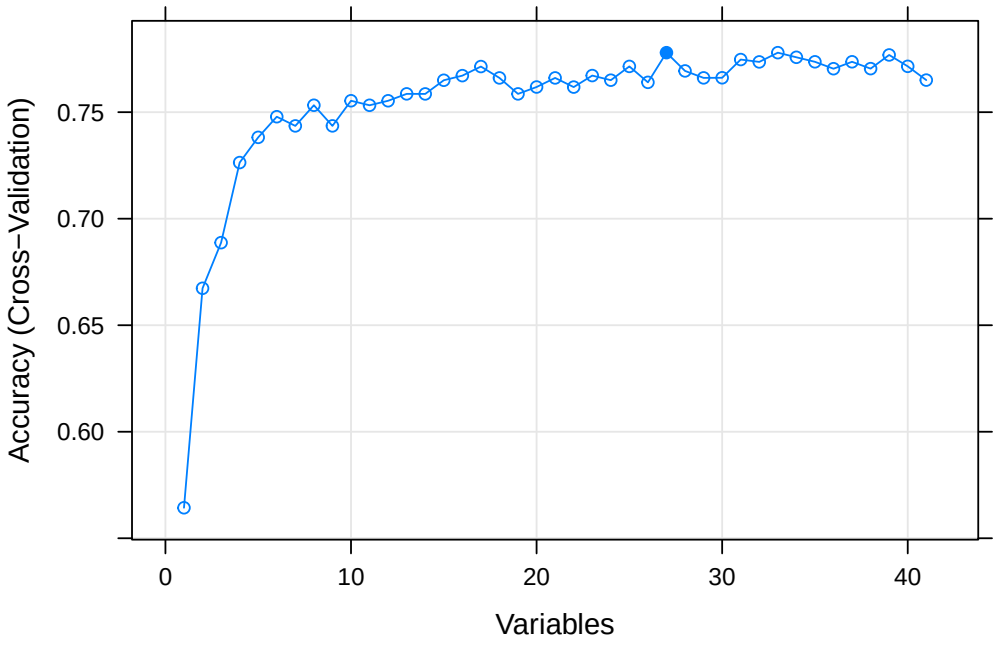


Figure 5.8. Results of recursive feature elimination.

moderators to focus on complaints that may require urgent attention. For only the highly severe cases (denoted with severity values 4 and 5 in the 5-point Likert scale), the classification accuracy is 85.3% in holdout set.

In addition to the system's performance, we can also inspect the trained system's internals and retrieve which features individually contribute most to the overall performance. In Figure 5.11, the top twenty most important features (from a total of 43 features) and their contributions are shown. Note that three of the five newly introduced inter-player communication features end up in this list. The communication features are potentially even more important, but they are difficult to analyze automatically (e.g. sarcasm is a potentially strong indicator, yet very difficult to detect). Bad language attempts, although caught by the system, are strong predictors of aggression. Suspect and accuser gender provide demographics, player profile features like games played, wins and rating show investment and involvement. Suspect words, silences and chat entries quantify social interaction and extraversion. Different games may have different features covering these factors.

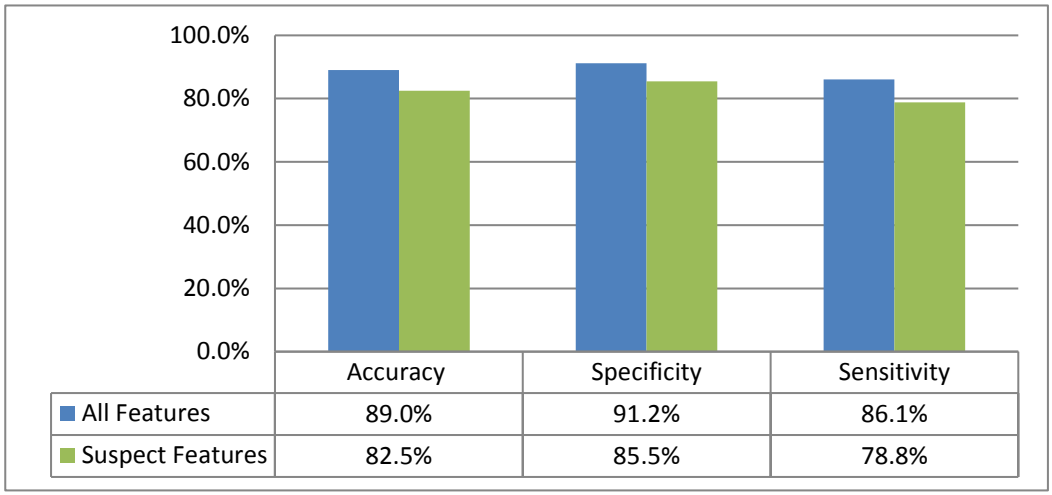


Figure 5.9. Results of GBM on training set with 5-fold cross validation.

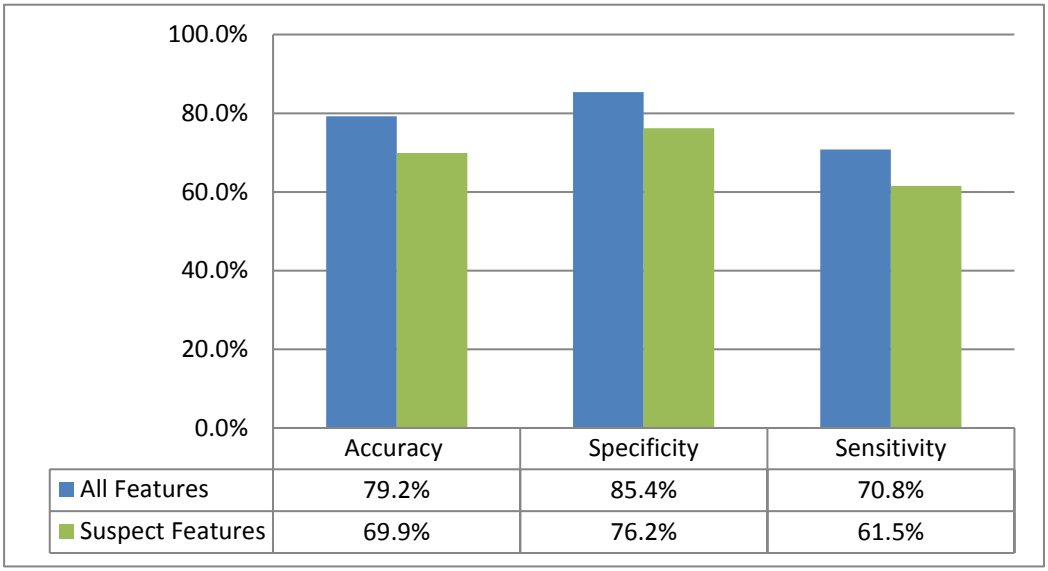


Figure 5.10. Results of GBM on the holdout set with 5-fold cross validation.

Most of the important features belong to suspect profile, while accusers also play some important role in classification of complaints. Since suspect-only cues are very important, just by looking at the profile of the suspect, as proposed in [18], yields good results. Using all features results in a statistically significant increase in accuracy (paired t-test, p value $p < 0.005$).

We have experimented with several other machine learning methods. On the holdout set, GBM performs better (79.2%) when compared with other methods we presented earlier as shown in 5.12.

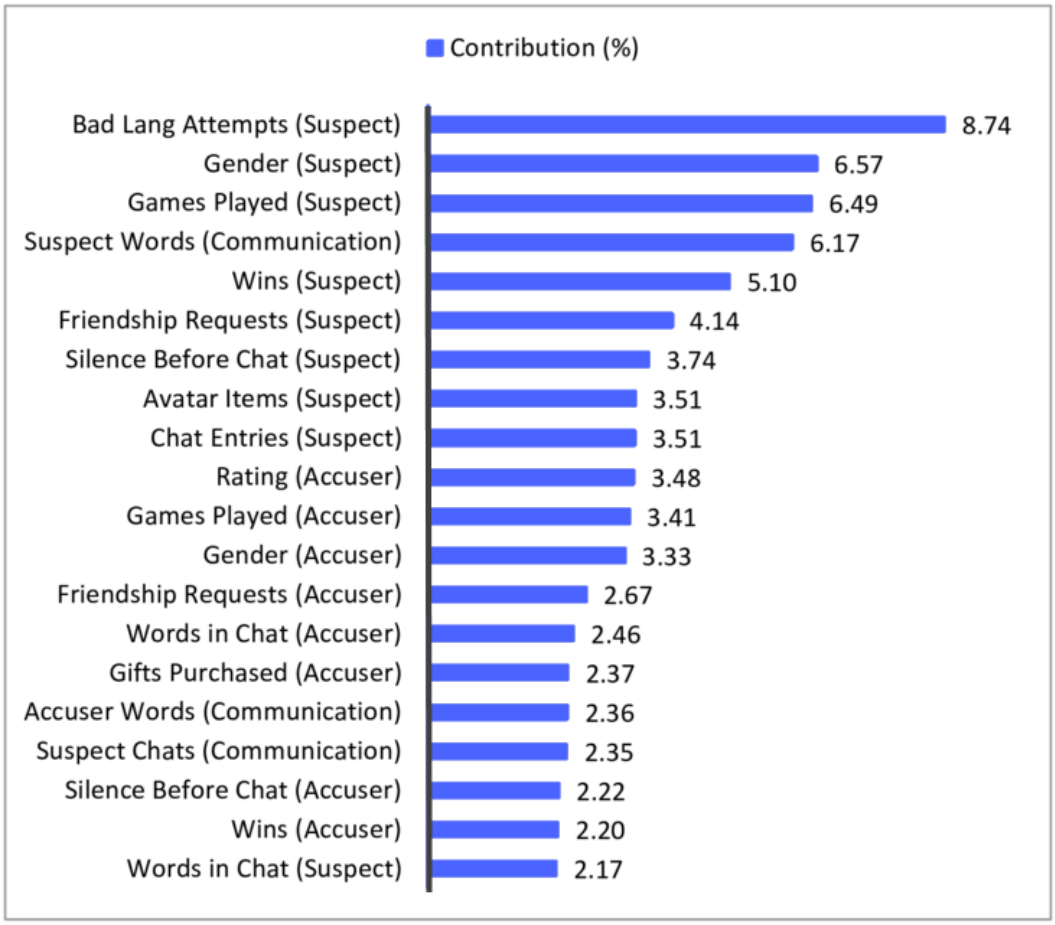


Figure 5.11. Top 20 most contributing features to prediction of genuine complaints.

We also report our early studies on classification of complaints with these machine learning methods by forming a feature vector consisting of both the suspect and the victim player profiles without introducing the communicative features. In this dataset, we have 932 complaint submissions in which both the victim and the suspect has played 5 or more games of Okey. We enforce this constraint in order to filter out those players who are new to the game and have not created a significant profile data yet. Although we did the experiments for feature subsets of gameplay, customer and community, the results are not satisfactory at all. So, here we only present the results when all features used for complaint classification in Figure 5.13.

Unfortunately, the results are not as good as we have anticipated in that setting. This can happen because of the issues with the dataset. For the cases of complaints with no genuine abusive behaviour (negative samples), there exists two different groups. One group involves complaints probably submitted for reasons such as mild aggression,

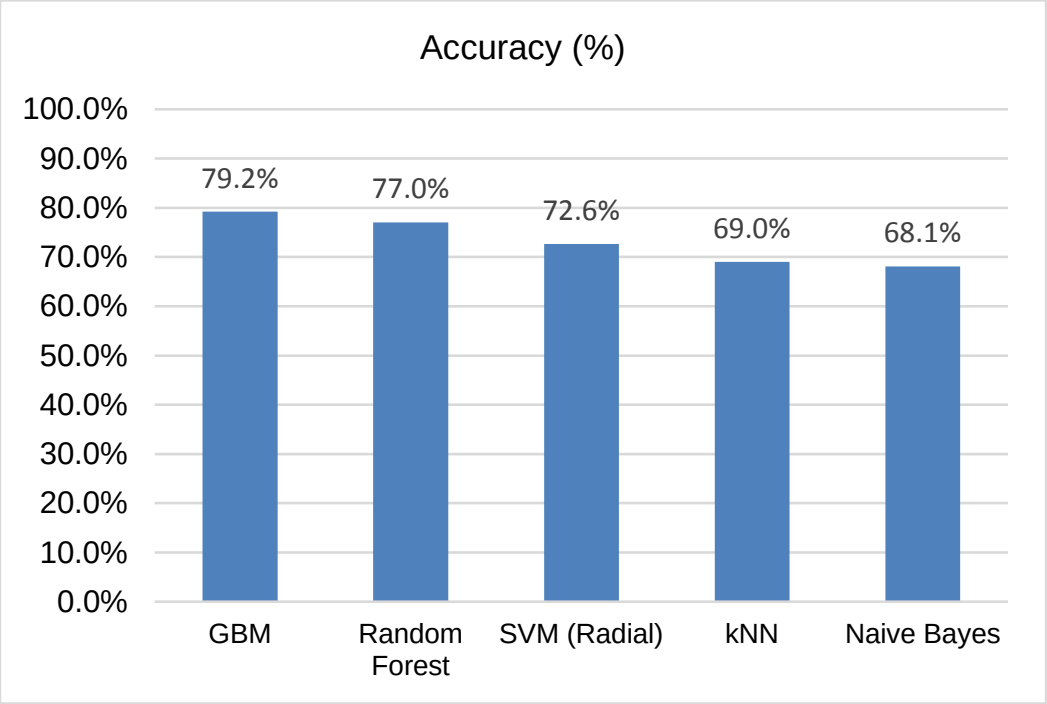


Figure 5.12. Performance of several machine learning methods on holdout set.

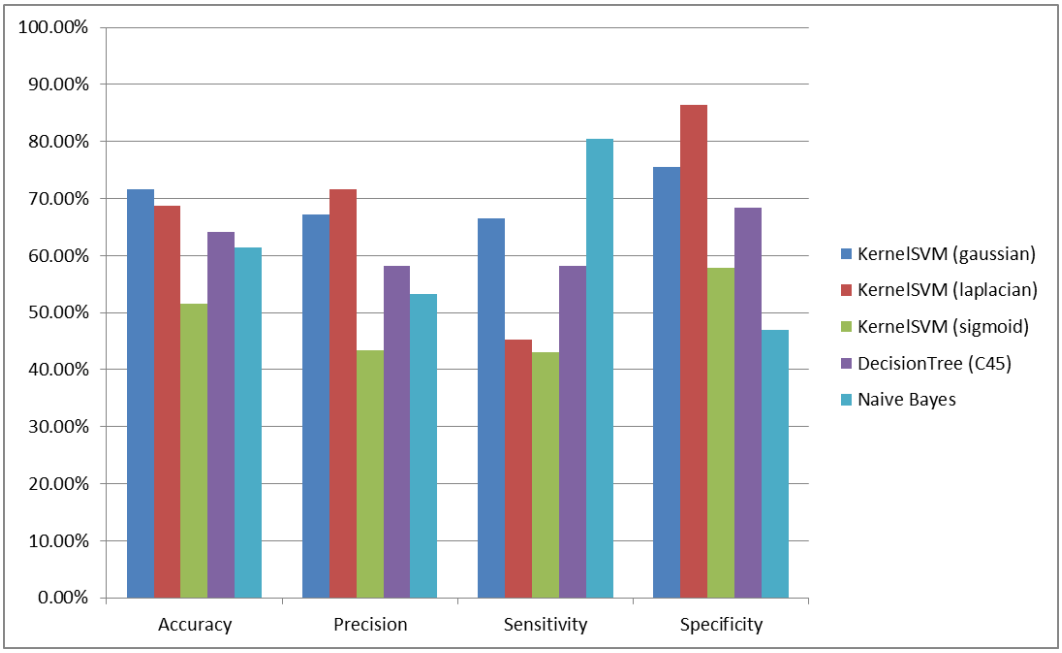


Figure 5.13. Results of complaint classification using suspect and victim features.

misunderstanding, etc. However, there are several abusive players who submit false complaints as well. These entries probably confuse the learning procedure.

Finally, we study victims in the complaints further. We repeated our classifica-

tion procedure done for suspects, this time to 250 victim players. We would like to see whether victims can be classified as accurate as suspects. Results for the victim classification is shown in Figure 5.14.

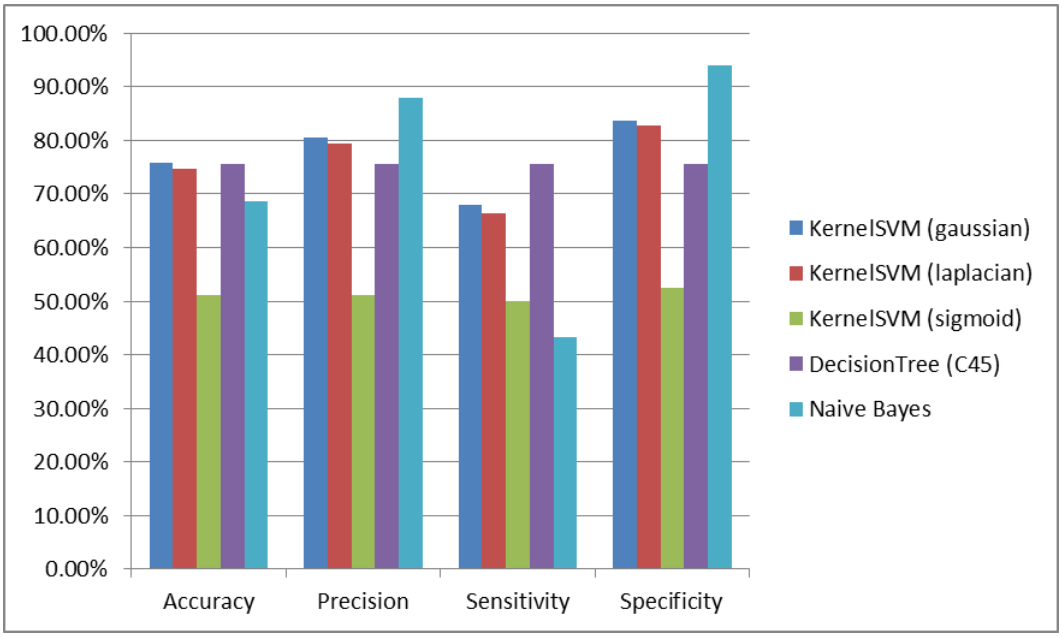


Figure 5.14. Results of victim classification.

Note that, the results are very comparable to that of abusive player classification.

6. Conclusions

6.1. Abusive Player Classification

In this study, we presented our analysis of player complaints data acquired from a real online social game. Our primary aim was to spot offenders by profiling players according to their in-game behavior and performance. We propose a feature vector for player profile and a binary clustering methodology, followed by an adaptive thresholding for confidence to classify players reported in complaints as genuine offenders. Our approach performs well enough to aid human moderators in terms of precision and sensitivity so that they can prioritize and schedule which complaints to focus on. By looking at the sensitivity scores, we observe that our methodology falls short of being able to classify all players in a fully automatized manner. Therefore, human moderators are still needed for precise judgments.

Our initial hypothesis was that player profiling and classification can be used as a common methodology to evaluate player behavior in online games. Our results confirm this claim, as we demonstrate a working system built on this premise.

We assess different types of features for detecting abuse automatically. In addition to features related to game performance and general player information, we propose social features that are game independent to quantify a player's social interaction within the virtual game community. As suspected, the social features contribute most to the automatic analysis of abusive behaviors, but player information (such as gender) and gaming behaviors (such as the regularity with which the game is played) are also highly informative.

The adaptation of the proposed method to different games would require the identification of features from the three feature types (i.e. gameplay, customer, community), and the provisioning of a training set of labelled instances. Since any major online game that involves social behavior requires a moderation support for player

complaints, our proposed approach can serve as a template to implement a tool for helping the moderation effort.

6.1.1. Limitations

We have performed our assessment on a single online gaming platform. While the factors we have investigated have their counterparts in most online social games, this is a major limitation of the study, as intensive usage of chat is not the case for all online social games. The players sometimes use actual voice over IP chat over other channels to improve game play experience, and for some games the interaction is limited to a set of signals (like emotes or gifts) that can be sent to other players.

For the database annotations, it was not possible to use multiple annotators. As stated earlier, the annotations are very time consuming for the expert moderator, since all the past interactions related to the complaint are manually processed. Consequently, annotations are expensive to obtain. On the other hand, our results on high-severity annotations demonstrate that the existing annotations are of high relevance. A very interesting and related platform we would like to investigate is the League of Legends online game, which has a tribunal system in which eligible players vote on complaint cases built from in-game reports and majority decisions lead to pardons or punishments[29] . Such community based decision-making may alleviate the need for costly annotations. Unfortunately, this system is not implemented in the Turkish edition of the game yet.

Despite the massive amount of data collected for this work, we do not have access to self-reports, surveys, and questionnaires. The nature of the problem mostly prevents the use of these tools. We plan to extend the game itself to present players reported as abusive with a questionnaire (e.g. Buss-Perry aggression questionnaire [132]) to be used as a benchmark, and to motivate the cooperation with in-game currency. Since such an addition to the game requires significant development and testing by the game developers, it is addressed as a future work. Furthermore, it is not certain whether

[29]League of Legends, http://na.leagueoflegends.com/tribunal/. [Online; accessed 18-March-2014].

cooperation of players can be ensured with this approach. In order to extend the analysis to preconditions of abuse, a more controlled experimental setup is required, where major factors like impulsivity and the players' level of education can be incorporated and assessed.

6.2. Complaint Classification

In this study, we have presented a methodology to semi-automatically identify genuine player complaints for verbal aggression and abusive behaviors in an online social game. Our method does not require human investigation and labeling of the complaints, nor annotation of chat messages during its operation, but since it is a supervised approach, the training of the system requires labeled data. Since these labels are already generated by the game moderators during their handling of day to day complaints, our approach does not require any labeling beyond what is normally performed in a gaming company.

Our previous study on player profiling suggested that using player data to spot abusive players is a valid approach [18]. With this study, we show that classifying complaints instead of players is better. In addition, we propose here the use of features from the accuser profile and communication features obtained from the interaction between both parties. An inspection of the individual contributions of the features confirmed our hypothesis. We have shown that the extended feature set outperforms the use of only suspected player profile.

The data driven nature of our approach makes it applicable to other online games, provided that a rich game-specific feature set is employed to model player profiles and communication profiles. We expect that the game implements a complaint mechanism, which will trigger the analysis. We also expect that a human moderator will be the final judge to select the appropriate course of action. The classification itself takes a short amount of time (less than a second), provided that the historical player data can be quickly accessed. Since the moderator response is on the order of minutes, if not hours, the approach is entirely scalable to the biggest online social games currently

in existence. We believe that online games which offer their players some means to socialize can benefit from our player profiling and interaction analysis in order to resolve conflicts among the players.

6.3. Extensions and Future Work

Our study on verbal affect analysis established that it is possible to normalize Turkish multiparty chat data to determine affective tones of the verbal interaction. We have obtained good results in the Valence dimension, which is arguably the most useful for abusive behavior analysis. Future work includes the integration of affective predictions as novel features to our complaint classification framework.

Abusive players who are banned from the system can create new users to join the system. Our investigation into chat biometrics showed that it is possible to identify people from their chat behavior to some extent. Especially, the word usage is revealing. Normalization impairs the results, which means that the errors and abbreviations used by players contain discriminative information. We see room for improvement on this topic as well and plan to explore the problem further.

Bibliography

1. Asteriadis, S., N. Shaker, K. Karpouzis, and G. N. Yannakakis, "Towards Player's Affective and Behavioral Visual Cues as drives to Game Adaptation", *LREC Workshop on Multimodal Corpora for Machine Learning, Istanbul*, 2012.

2. van Lankveld, G., P. Spronck, J. van den Herik, and A. Arntz, "Games as personality profiling tools", *Computational Intelligence and Games (CIG), 2011 IEEE Conference on*, pp. 197–202, IEEE, 2011.

3. Egenfeldt-Nielsen, S., J. H. Smith, and S. P. Tosca, *Understanding video games: The essential introduction*, Routledge, 2013.

4. Griffiths, M., "Violent video games and aggression: A review of the literature", *Aggression and Violent Behavior*, Vol. 4, No. 2, pp. 203–212, 1999.

5. Ferguson, C. J., "Violent video games and the supreme court: lessons for the scientific community in the wake of Brown v. Entertainment Merchants Association.", *American Psychologist*, Vol. 68, No. 2, p. 57, 2013.

6. Uthus, D. C. and D. W. Aha, "Multiparticipant chat analysis: A survey", *Artificial Intelligence*, Vol. 199-–200, pp. 106–121, 2013.

7. Reynolds, K., A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying", *Proc. IEEE ICMLA*, pp. 241–244, 2011.

8. Vinciarelli, A. and G. Mohammadi, "A Survey of Personality Computing", *IEEE Trans. on Affective Computing*, Vol. 5, No. 3, pp. 273–291, 2014.

9. Cristani, M., A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis", *Proc. ACM MM*, pp. 213–222, ACM, 2013.

10. Brunswik, E., *Perception and the representative design of psychological experiments*, Univ of California Press, 1956.

11. Yudofsky, S., J. Silver, W. Jackson, J. Endicott, and D. Williams, "The Overt Aggression Scale for the objective rating of verbal and physical aggression", *American journal of psychiatry*, Vol. 143, No. 1, pp. 35–39, 1986.

12. Archer, J., G. Kilpatrick, and R. Bramwell, "Comparison of two aggression inventories", *Aggressive Behavior*, Vol. 21, No. 5, pp. 371–380, 2006.

13. Finkelhor, D., "Commentary: Cause for alarm? Youth and internet risk research–a commentary on Livingstone and Smith (2014)", *Journal of child psychology and psychiatry*, Vol. 55, No. 6, pp. 655–658, 2014.

14. France, K., A. Danesh, and S. Jirard, "Informing aggression–prevention efforts by comparing perpetrators of brief vs. extended cyber aggression", *Computers in Human Behavior*, Vol. 29, No. 6, pp. 2143–2149, 2013.

15. Li, Q., "New bottle but old wine: A research of cyberbullying in schools", *Computers in Human Behavior*, Vol. 23, No. 4, pp. 1777–1791, 2007.

16. Kwan, G. C. E. and M. M. Skoric, "Facebook bullying: An extension of battles in school", *Computers in Human Behavior*, Vol. 29, No. 1, pp. 16 – 25, 2013.

17. Balci, K. and A. A. Salah, "Player profiling and offender classification from player complaints in online social games", *Workshop on Design and Evaluation of Sociability in Online Games at CHI, Paris*, 2013.

18. Balci, K. and A. A. Salah, "Automatic analysis and identification of verbal aggression and abusive behaviors for online social games", *Computers in Human Behavior*, Vol. 53, pp. 517–526, 2015.

19. Balci, K. and A. A. Salah, "Automatic classification of player complaints in social games", *Computational Intelligence and AI in Games, IEEE Transactions on*, ,

No. 99, pp. 1–1, 2015.

20. Aydın Oktay, E., K. Balcı, and A. A. Salah, "Automatic Assessment of Dimensional Affective Content in Turkish Multi-party Chat Messages", *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, pp. 19–24, ACM, 2015.

21. Aichner, T. and F. Jacob, "Measuring the degree of corporate social media use", *International Journal of Market Research*, Vol. 57, No. 2, pp. 257–275, 2015.

22. Gonçalves, B. and J. J. Ramasco, "Human dynamics revealed through Web analytics", *Physical Review E*, Vol. 78, No. 2, p. 026123, 2008.

23. McAfee, A., E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data", *The management revolution. Harvard Bus Rev*, Vol. 90, No. 10, pp. 61–67, 2012.

24. Menzies, T. and T. Zimmermann, "Software analytics: so what?", *Software, IEEE*, Vol. 30, No. 4, pp. 31–37, 2013.

25. Davenport, T. H., *Enterprise analytics: Optimize performance, process, and decisions through big data*, Pearson Education, 2013.

26. Kohavi, R., R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web: listen to your customers not to the hippo", *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 959–967, ACM, 2007.

27. Fisher, D., R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics", *interactions*, Vol. 19, No. 3, pp. 50–59, 2012.

28. Fields, T. V., "Game Industry Metrics Terminology and Analytics Case Study", Seif El-Nasr, M., A. Drachen, and A. Canossa (editors), *Game Analytics*, pp. 53–71, Springer London, 2013.

29. Zimmermann, T. and N. Nagappan, "Software Analytics for Digital Games.", *Software Engineering*, pp. 23–24, 2014.

30. Couper, M. P., "Review: Web surveys: A review of issues and approaches", *Public opinion quarterly*, pp. 464–494, 2000.

31. Lucas, S. M., M. Mateas, M. Preuss, P. Spronck, and J. Togelius, "Artificial and computational intelligence in games (Dagstuhl Seminar 12191)", *Dagstuhl Reports*, Vol. 2, No. 5, 2012.

32. Yannakakis, G. and J. Togelius, "A Panorama of Artificial and Computational Intelligence in Games", *IEEE Trans. on Computational Intelligence and AI in Games*, 2015.

33. Yannakakis, G. N., P. Spronck, D. Loiacono, and E. André, "Player modeling", *Dagstuhl Follow-Ups*, Vol. 6, 2013.

34. Dumais, S., R. Jeffries, D. M. Russell, D. Tang, and J. Teevan, "Understanding User Behavior Through Log Data and Analysis", Olson, J. S. and W. A. Kellogg (editors), *Ways of Knowing in HCI*, pp. 349–372, Springer, 2014.

35. Loh, C. S. and Y. Sheng, "Measuring Expert Performance for Serious Games Analytics: From Data to Insights", Loh, C. S., Y. Sheng, and D. Ifenthaler (editors), *Serious Games Analytics*, Advances in Game-Based Learning, pp. 101–134, Springer, 2015.

36. Schouten, B., R. Tieben, A. van de Ven, and D. Schouten, "Human behavior analysis in ambient gaming and playful interaction", Salah, A. A. and T. Gevers (editors), *Computer Analysis of Human Behavior*, pp. 387–403, Springer, 2011.

37. Yan, J. and H. Choi, "Security issues in online games", *The Electronic Library*, Vol. 20, No. 2, pp. 125–133, 2002.

38. Shim, K., R. Sharan, and J. Srivastava, "Player performance prediction in mas-

sively multiplayer online role-playing games (MMORPGS)", *Advances in Knowledge Discovery and Data Mining*, pp. 71–80, 2010.

39. Mellon, L., "Applying metrics driven development to MMO costs and risks", *Versant Corporation, Technical Report*, pp. 1–9, 2009.

40. Xie, H., D. Kudenko, S. Devlin, and P. Cowling, "Predicting Player Disengagement in Online Games", Cazenave, T., M. Winands, and Y. Björnsson (editors), *Computer Games*, pp. 133–149, Springer, 2014.

41. Bauckhage, C., A. Drachen, and R. Sifa, "Clustering Game Behavior Data", *IEEE Trans. on Computational Intelligence and AI in Games*, 2014.

42. Hadiji, F., R. Sifa, A. Drachen, C. Thurau, K. Kersting, and C. Bauckhage, "Predicting player churn in the wild", *Proc. IEEE CIG*, 2014.

43. Drachen, A., C. Thurau, R. Sifa, and C. Bauckhage, "A comparison of methods for player clustering via behavioral telemetry", *arXiv preprint arXiv:1407.3950*, 2014.

44. Drachen, A., A. Canossa, and G. N. Yannakakis, "Player modeling using self-organization in Tomb Raider: Underworld", *Proc. IEEE CIG*, 2009.

45. Yee, N., N. Ducheneaut, L. Nelson, and P. Likarish, "Introverted elves & conscientious gnomes: the expression of personality in World of Warcraft", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 753–762, ACM, 2011.

46. Bean, A. and G. Groth-Marnat, "Video Gamers and Personality: A Five-Factor Model to Understand Game Playing Style.", 2014.

47. John, O. P., E. M. Donahue, and R. L. Kentle, "The big five inventory—versions 4a and 54", *Berkeley: University of California, Berkeley, Institute of Personality and Social Research*, 1991.

48. Markey, P. M. and C. N. Markey, "Vulnerability to violent video games: a review and integration of personality research.", *Review of General Psychology*, Vol. 14, No. 2, p. 82, 2010.

49. Miller, J. L. and J. Crowcroft, "Avatar movement in World of Warcraft battlegrounds", *Proceedings of the 8th annual workshop on Network and systems support for games*, p. 1, IEEE Press, 2009.

50. Lewis, C. and N. Wardrip-Fruin, "Mining game statistics from web services: a World of Warcraft armory case study", *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, pp. 100–107, ACM, 2010.

51. Ashton, M. and C. Verbrugge, "Measuring cooperative gameplay pacing in World of Warcraft", *Proceedings of the 6th International Conference on Foundations of Digital Games*, pp. 77–83, ACM, 2011.

52. Digman, J. M., "Personality structure: Emergence of the five-factor model", *Annual Review of Psychology*, Vol. 41, No. 1, pp. 417–440, 1990.

53. Huang, J., T. Zimmermann, N. Nagapan, C. Harrison, and B. C. Phillips, "Mastering the art of war: how patterns of gameplay influence skill in Halo", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 695–704, ACM, 2013.

54. Hullett, K., N. Nagappan, E. Schuh, and J. Hopson, "Empirical analysis of user data in game software development", *Empirical Software Engineering and Measurement (ESEM), 2012 ACM-IEEE International Symposium on*, pp. 89–98, IEEE, 2012.

55. Weber, B. G. and M. Mateas, "A data mining approach to strategy prediction", *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pp. 140–147, IEEE, 2009.

56. Weber, B. G., M. John, M. Mateas, and A. Jhala, "Modeling Player Retention in Madden NFL 11.", *IAAI*, 2011.

57. Yan, E. Q., J. Huang, and G. K. Cheung, "Masters of Control: Behavioral Patterns of Simultaneous Unit Group Manipulation in StarCraft 2", *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3711–3720, ACM, 2015.

58. Suznjevic, M. and M. Matijasevic, "Player behavior and traffic characterization for MMORPGs: a survey", *Multimedia systems*, Vol. 19, No. 3, pp. 199–220, 2013.

59. Wallner, G. and S. Kriglstein, "Visualization-based analysis of gameplay data–a review of literature", *Entertainment Computing*, Vol. 4, No. 3, pp. 143–155, 2013.

60. Lewis, B., *Istanbul and the Civilization of the Ottoman Empire*, University of Oklahoma Press, 1963.

61. Elo, A. E., *The rating of chessplayers, past and present*, Vol. 3, Batsford London, 1978.

62. Knapp, M. L., *Nonverbal communication in human interaction*, Cengage Learning, 2012.

63. Aran, O. and D. Gatica-Perez, "Analysis of group conversations: modeling social verticality", Salah, A. A. and T. Gevers (editors), *Computer Analysis of Human Behavior*, pp. 293–322, Springer, 2011.

64. Kreyszig, E., *Advanced engineering mathematics*, Wiley New York, 1979.

65. Zeng, Z., M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 31, No. 1, pp. 39–58, 2009.

66. Salah, A. A., N. Sebe, and T. Gevers, "Communication and automatic interpretation of affect from facial expressions", *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives. IGI Global (to appear)*, 2010.

67. Zhang, L. and J. Barnden, "Affect sensing using linguistic, semantic and cognitive cues in multi-threaded improvisational dialogue", *Cognitive Computation*, Vol. 4, No. 4, pp. 436–459, 2012.

68. Sánchez, J. A., N. P. Hernández, J. C. Penagos, and Y. Ostróvskaya, "Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states", *Proceedings of VII Brazilian symposium on Human factors in computing systems*, pp. 66–72, ACM, 2006.

69. Tsetserukou, D., A. Neviarouskaya, H. Prendinger, N. Kawakami, M. Ishizuka, and S. Tachi, "iFeel_IM! Emotion enhancing garment for communication in affect sensitive instant messenger", *Human Interface and the Management of Information. Designing Information Environments*, pp. 628–637, 2009.

70. Tat, A. and S. Carpendale, "CrystalChat: Visualizing personal chat history", *Proc. 39th Annual Hawaii Int. Conf. on System Sciences*, Vol. 3, p. 58c, IEEE, 2006.

71. Zheng, D., F. Tian, J. Liu, Q. Zheng, and J. Qin, "Emotion Chat: A Web Chatroom with Emotion Regulation for E-Learners", *Physics Procedia*, Vol. 25, pp. 763–770, 2012.

72. Shin, H., J. Lee, J. Park, Y. Kim, H. Oh, and T. Lee, "A tactile emotional interface for instant messenger chat", *Human Interface and the Management of Information. Interacting in Information Environments*, pp. 166–175, Springer, 2007.

73. Wang, H., H. Prendinger, and T. Igarashi, "Communicating emotions in online

chat using physiological sensors and animated text", *CHI'04 extended abstracts on Human factors in computing systems*, pp. 1171–1174, ACM, 2004.

74. Kolz, B., J. M. Garrido, and Y. Laplaza, "Automatic prediction of emotions from text in Spanish for expressive speech synthesis in the chat domain", *Procesamiento del Lenguaje Natural*, Vol. 52, pp. 61–68, 2014.

75. Pang, B. and L. Lee, "Opinion mining and sentiment analysis", *Foundations and trends in information retrieval*, Vol. 2, No. 1-2, pp. 1–135, 2008.

76. Ma, C., H. Prendinger, and M. Ishizuka, "Emotion estimation and reasoning based on affective textual interaction", *Affective computing and intelligent interaction*, pp. 622–628, Springer, 2005.

77. Dey, L., M.-U. Asad, N. Afroz, and R. P. D. Nath, "Emotion extraction from real time chat messenger", *Proc. ICIEV*, pp. 1–5, IEEE, 2014.

78. Neviarouskaya, A., H. Prendinger, and M. Ishizuka, "Affect analysis model: novel rule-based approach to affect sensing from text", *Natural Language Engineering*, Vol. 17, No. 01, pp. 95–135, 2011.

79. Brooks, M., K. Kuksenok, M. K. Torkildson, D. Perry, J. J. Robinson, T. J. Scott, O. Anicello, A. Zukowski, P. Harris, and C. R. Aragon, "Statistical affect detection in collaborative chat", *Proc. CSCW*, pp. 317–328, ACM, 2013.

80. Liu, H., H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge", *Proc. IUI*, pp. 125–132, ACM, 2003.

81. Lenat, D. B., "CYC: A large-scale investment in knowledge infrastructure", *Communications of the ACM*, Vol. 38, No. 11, pp. 33–38, 1995.

82. Cakmak, O., A. Kazemzadeh, D. Can, S. Yildirim, and S. Narayanan, "Root-word analysis of Turkish emotional language", *Corpora for Research on Emotion Sentiment & Social Signals*, 2012.

83. Boynukalin, Z. and P. Karagoz, "Emotion Analysis on Turkish Texts", Gelenbe, E. and R. Lent (editors), *Information Sciences and Systems*, Vol. 264 of *LNEE*, pp. 159–168, 2013, http://dx.doi.org/10.1007/978-3-319-01604-7_16.

84. Yıldırım, E., F. S. Çetin, G. Eryiğit, and T. Temel, "The Impact of NLP on Turkish Sentiment Analysis", *Proc. of the Int. Conf. on Turkic Language Processing*, pp. 7–13, 2014.

85. Vural, A. G., B. B. Cambazoglu, P. Senkul, and Z. O. Tokgoz, "A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish", *Computer and Information Sciences III*, pp. 437–445, Springer, 2013.

86. Dehkharghani, R., Y. Saygin, B. Yanikoglu, and K. Oflazer, "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis", *Language Resources and Evaluation*, pp. 1–19, 2015, http://dx.doi.org/10.1007/s10579-015-9307-6.

87. Bradley, M. M. and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings", Technical report, C-1, The Center for Research in Psychophysiology, Univ. of Florida, 1999.

88. Bradley, M. M. and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential", *Journal of behavior therapy and experimental psychiatry*, Vol. 25, No. 1, pp. 49–59, 1994.

89. Sheng, V. S., F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers", *Proc. 14th ACM SIGKDD*, pp. 614–622, 2008.

90. Warriner, A. B., V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas", *Behavior research methods*, Vol. 45, No. 4, pp. 1191–1207, 2013.

91. Torunoğlu, D. and G. Eryiğit, "A Cascaded Approach for Social Media Text

Normalization of Turkish", *5th Workshop on Language Analysis for Social Media (LASM) at EACL*, pp. 62–70, Association for Computational Linguistics, Gothenburg, Sweden, April 2014.

92. Oflazer, K., "Two-level description of Turkish morphology", *Literary and linguistic computing*, Vol. 9, No. 2, pp. 137–148, 1994.

93. Oflazer, K., E. Göçmen, and C. Bozsahin, "An Outline of Turkish Morphology", *Report on Turkish Natural Language Processing Initiative Project*, 1994.

94. D'Mello, S. K. and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems", *ACM Computing Surveys (CSUR)*, Vol. 47, No. 3, p. Article 43, 2015.

95. Ali, N., M. Price, and R. Yampolskiy, "BLN-Gram-TF-ITF as a new Feature for Authorship Identification", *Academy of Science and Engineering (ASE) BIG-DATA/SOCIALCOM/CYBERSECURITY Conference*, 2014.

96. Stamatatos, E., "A survey of modern authorship attribution methods", *Journal of the American Society for information Science and Technology*, Vol. 60, No. 3, pp. 538–556, 2009.

97. Šarkute, L. and A. Utka, "The Effect of Author Set Size in Authorship Attribution for Lithuanian", *Nordic Conference of Computational Linguistics NODALIDA 2015*, p. 87, 2015.

98. Zheng, R., J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques", *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 3, pp. 378–393, 2006.

99. Inches, G., M. Harvey, and F. Crestani, "Finding participants in a chat: Authorship attribution for conversational documents", *Social Computing (SocialCom),*

2013 International Conference on, pp. 272–279, IEEE, 2013.

100. Layton, R., S. McCombie, and P. Watters, "Authorship attribution of IRC messages using inverse author frequency", *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, pp. 7–13, IEEE, 2012.

101. Juola, P., "Authorship attribution for electronic documents", *Advances in digital forensics II*, pp. 119–130, Springer, 2006.

102. Sanderson, C. and S. Guenter, "On authorship attribution via Markov chains and sequence kernels", *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, Vol. 3, pp. 437–440, IEEE.

103. Kešelj, V., F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution", *Proceedings of the conference pacific association for computational linguistics, PACLING*, Vol. 3, pp. 255–264, 2003.

104. Mikros, G. K. and K. Perifanos, "Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles.", *AAAI Spring Symposium: Analyzing Microtext*, 2013.

105. Tufan, T. and A. K. Görür, "Author Identification for Turkish Texts", *Cankaya University Journal of Arts and Sciences*, Vol. 1, No. 7, 2007.

106. Amasyalı, M. F. and B. Diri, "Automatic Turkish text categorization in terms of author, genre and gender", *Natural Language Processing and Information Systems*, pp. 221–226, Springer, 2006.

107. Kucukyilmaz, T., B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication", *Information Processing & Management*, Vol. 44, No. 4, pp. 1448–1466, 2008.

108. Layton, R., P. Watters, and R. Dazeley, "Recentred local profiles for authorship attribution", *Natural Language Engineering*, Vol. 18, No. 03, pp. 293–312, 2012.

109. Eryigit, G., "ITU Turkish NLP web service", *EACL 2014*, p. 1, 2014.

110. Vapnik, V., *The Nature of Statistical Learning Theory*, Springer Verlag, 2000.

111. Aizerman, A., E. M. Braverman, and L. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning", *Automation and remote control*, Vol. 25, pp. 821–837, 1964.

112. Platt, J. *et al.*, "Sequential minimal optimization: A fast algorithm for training support vector machines", 1998.

113. Quinlan, J. R., "Induction of decision trees", *Machine learning*, Vol. 1, No. 1, pp. 81–106, 1986.

114. Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

115. Mitchell, T. M., "Machine learning. 1997", *Burr Ridge, IL: McGraw Hill*, Vol. 45, 1997.

116. Hyafil, L. and R. L. Rivest, "Constructing optimal binary decision trees is NP-complete", *Information Processing Letters*, Vol. 5, No. 1, pp. 15–17, 1976.

117. Zhang, H., "The optimality of naive Bayes", *AA*, Vol. 1, No. 2, p. 3, 2004.

118. Mladenic, D. and M. Grobelnik, "Feature selection for unbalanced class distribution and Naive Bayes", *In Proceedings of the 16th International Conference on Machine Learning (ICML*, pp. 258–267, Morgan Kaufmann Publishers, 1999.

119. Maloof, M. A., "Learning when data sets are imbalanced and when costs are unequal and unknown", *ICML-2003 workshop on learning from imbalanced data sets II*, Vol. 2, pp. 2–1, 2003.

120. Caruana, R. and A. Niculescu-mizil, "An Empirical Comparison of Supervised Learning Algorithms", *In Proc. 23 rd Intl. Conf. Machine learning (ICML'06)*,

pp. 161–168, 2006.

121. Herbrich, R., T. Graepel, and C. Campbell, "Bayes point machines: Estimating the Bayes point in kernel space", *IJCAI Workshop SVMs*, pp. 23–27, 1999.

122. Herbrich, R., T. Graepel, and C. Campbell, "Bayes point machines", *The Journal of Machine Learning Research*, Vol. 1, pp. 245–279, 2001.

123. Hearst, M., S. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines", *Intelligent Systems and their Applications, IEEE*, Vol. 13, No. 4, pp. 18–28, 1998.

124. Chang, E., K. Goh, G. Sychay, and G. Wu, "CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines", *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 13, No. 1, pp. 26–38, 2003.

125. Friedman, J. H., "Greedy function approximation: a gradient boosting machine", *Annals of Statistics*, pp. 1189–1232, 2001.

126. Natekin, A. and A. Knoll, "Gradient boosting machines, a tutorial", *Frontiers in Neurorobotics*, Vol. 7, No. 21, pp. 1–21, 2013.

127. Kuhn, M., "Building predictive models in R using the caret package", *Journal of Statistical Software*, Vol. 28, No. 5, pp. 1–26, 2008.

128. Minka, T., J. Winn, J. Guiver, and D. Knowles, "Infer.NET 2.5", 2012, microsoft Research Cambridge. http://research.microsoft.com/infernet.

129. Minka, T. P., "Expectation propagation for approximate Bayesian inference", *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 362–369, Morgan Kaufmann Publishers Inc., 2001.

130. Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

131. Granitto, P. M., C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products", *Chemometrics and Intelligent Laboratory Systems*, Vol. 83, No. 2, pp. 83–90, 2006.

132. Buss, A. H. and M. Perry, "The aggression questionnaire", *Journal of Personality and Social Psychology*, Vol. 63, No. 3, pp. 452–459, 1992.