Abordagens Inteligentes para Identificação de Toxicidade em Jogos MOBA*

Alexandre Luis Batista da Silva, Alexandre Donnelly Vaz, Farmy Gonçalves Ferreira da Silva, Lincoln Magalhães Costa
Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
absilva@cos.ufrj.br, adonnellyvaz@cos.ufrj.br, farmygfs@cos.ufrj.br, costa@cos.ufrj,br

***Resumo*—De acordo com New Zoo [1], jogos Multiplayer Online Battle Arena (MOBA) estão no topo da lista de gêneros mais jogados, com títulos como League of Legends (LoL) e Defense of The Ancients (DOTA). Mora-Cantallops [2] exibe em seu estudo que um grande esforço tem sido dedicado para a análise de comportamentos tóxicos nestes jogos. Neste sentido, este trabalho apresenta um modelo de aprendizagem supervisionada para classificação de mensagens tóxicas em DOTA 2. Os resultados obtidos demonstram um acurácia de 87% na identificação de mensagens tóxicas.**

***Index Terms*—Toxicidade em jogos, Jogos MOBA, Aprendizagem Supervisionada, Processamento de Linguagem Natural, Toxicidade, DOTA2.**

I. Introdução

Os jogos digitais representaram um mercado de 120,1 bilhões de dólares no ano de 2019 [3], grande parte desse mercado é dos jogos digitais denominados de esporte eletrônico. Uma característica comum nos projetos de jogos digitais é permitir a interação entre jogadores através da utilização de comunicação por ícones, bate papos por áudio e texto. Essa possibilidade de socialização utilizada nos jogos digitais é ao mesmo tempo um grande atrativo e também um grande problema para a comunidade do jogo.Há diversos relatos sobre a utilização ruim dos meios de comunicação em jogos digitais [4]–[7], que é comumente referenciada como toxicidade. Este abuso de comunicação é identificado por uma série de expressões de agressividade e ofensas destinadas a outros jogadores.

A indústria de jogos busca diversas formas de minimizar estes efeitos negativos em seus jogos digitais, desde sistemas de denúncias para análise posterior de uma equipe de especialistas, como a busca por um sistema mais automatizado para detecção de mensagens inadequadas. Sem a presença dessas formas de proteção os jogadores podem se sentir resignados, e independente do resultado do jogo, podem se arrepender da decisão de ter jogado por conta do sentimento de abuso ou ofensa. Essas situações são comuns em jogos de Multiplayer Online Battle Arena (MOBA), que hoje movimentam as maiores premiações e atrativos para jogadores.

Um MOBA é caracterizado por uma dinâmica altamente competitiva onde jogadores são divididos em um modelo de n vs n jogadores. As duas maiores franquias realizam jogos de 5 vs 5 jogadores, onde a comunicação pode ocorrer entre a equipe, isto é, entre os cinco jogadores ou entre todos os dez participantes do jogos. A hostilidade da comunicação nessas partidas e o desafio que é identificá-las torna a toxicidade em jogos um tema importante para o desenvolvimento de jogos online.

O desafio, de forma mais objetiva, é que as mensagens de jogos multijogadores geralmente são ricas em erros de escrita, abreviações, expressões, vocabulário do jogo, e frases fora de contexto. Isto dificulta o reconhecimento de algum tipo de padrão que caracterize mensagens tóxicas. Este trabalho representa uma tentativa de identificar a toxicidade textual presente em jogos multijogadores, através da interação entre os jogadores nos chats das partidas. A toxicidade foi tratada de forma binária, portanto, o sistema proposto consiste em classificar cada mensagem enviada como tóxica ou não-tóxica.E para a realização deste estudo, utilizou-se a base de dados do jogo DOTA 2.

II. Metodologia

A. Origem dos Dados

O jogo Defense of the Ancients 2, também conhecido como DOTA 2, fornece uma API (*Application Programming Interface*) em código aberto chamada de OpenDota [8] que coleta dados de diversas partidas dos jogadores que deram permissão na forma de opt-in, e de partidas profissionais e competitivas automaticamente. Está API fornece um serviço de requisições no padrão REST para que as informações coletadas sejam disponibilizadas a comunidade. Embora esta opção seja ótima para um estudo mais profundo, a base de dados utilizada é proveniente do trabalho de Joe Ramir disponibilizada na plataforma Kaggle [9] denominada ”Dota 2 matches Dataset” [10] . Esta base de dados foi retirada também da plataforma OpenDota e estruturada com diversas informações sobre as partidas, jogadores, equipes, personagens e outras informações.

A base de dados usada para o estudo inclui várias colunas para cada mensagem, além do texto da mensagem em si, entre os atributos estão: o número de identificação da partida, o usuário do jogador que mandou a mensagem e o horário da mensagem relativo ao começo do jogo.

B. Domínio dos Especialistas

Os especialistas deste artigo possuem um nível variado de familiaridade com o DOTA e jogos MOBA em geral. Enquanto alguns são jogadores assíduos do DOTA, League of Legends e outros títulos do gênero, outros tem conhecimento de tais jogos apenas por relatos de terceiros e conteúdo online. Essa diferença de familiaridade teve influência na classificação de mensagens tóxicas, com uma curva de aprendizado quanto aos termos particulares ao jogo. Um exemplo é o uso da palavra *creep*, que em um contexto normal da língua inglesa se traduz como "alguém esquisito" e tem uma conotação negativa, mas no DOTA é um tipo de componente, que inclusive é jogável, portanto tem uma conotação neutra. Esse exemplo mostra que termos particulares ao DOTA ou a jogos MOBA em geral, seja de componentes, estratégias, objetivos, ambientes ou informações, fazem com que algumas mensagens tenham o seu nível de toxicidade dependente do contexto.

C. Seleção, Limpeza e Classificação dos Dados

Os dados foram selecionados com os seguintes critérios: mensagens de texto que estivessem em inglês, espanhol e português foram mantidos, demais idiomas foram ignorados. Mensagens com *emoji*, apenas números, ou símbolos especiais, isto é, @, * e similares, também foram ignorados. Ao todo foram ignoradas cerca de 1750 mensagens. Baseado nestes critérios, um total de 8780 mensagens foram compartilhadas entre os especialistas para a realização de uma classificação binária simples, onde as mensagens consideradas tóxicas foram classificadas com um (1) e as não tóxicas com zero (0). Um total de 1903 partidas tiveram suas mensagens classificadas por este trabalho.

D. Modelo de Aprendizagem e Classificação

Após a classificação e limpeza dos dados, descrita na subseção anterior, os dados refinados foram postos como entrada para o modelo de aprendizado desse estudo, que utilizou a ferramenta KNIME com o auxílio de nós do Palladian. O KNIME é um programa de ciência de dados no qual um *pipeline* de processos é formado para executar tarefas, de ponta-a-ponta, nas áreas de mineração de dados e aprendizado de máquina, entre outras. O processo inicia com a leitura dos dados classificados e com indicação das mensagens inclassificáveis e passa para um nó que particiona os dados para a validação cruzada e inicia um loop no qual cada iteração é designada para uma das partições da validação [11]. O nó de validação cruzada separa uma parcela dos dados para o aprendizado, feito pelo *Text Classifier Learner*, que recebe o arquivo com os dados manualmente pré-classificados pelos especialistas e usa uma tabela com pesos para cada termo para determinar o quão provável cada característica é para as categorias [12]. Em sequência vem o *Text Classifier Predictor*, que recebe o modelo estabelecido pelo *Text Classifier Learner* e os dados de testes particionados pelo *X-Partitioner* para classificar a toxicidade das mensagens e medir a eficácia do modelo. Por último, os dados passam pelo *X-Aggregator* que avalia a eficácia do modelo nos dados como um todo [13] e passa esses resultados para o nó *Scorer*. Este nó compara duas colunas por seus pares de valor de atributo e mostra a matriz de confusão, ou seja, quantas linhas de cada atributo e sua classificação correspondem. A saída do nó é a matriz de confusão com o número de correspondências em cada célula. Além disso, a segunda porta de saída relata uma série de estatísticas de precisão como verdadeiros positivos, falsos positivos, verdadeiros negativos, falsos negativos, recuperação, precisão, sensibilidade, especificidade, medida F, bem como a precisão geral e o coeficiente Kappa [14].

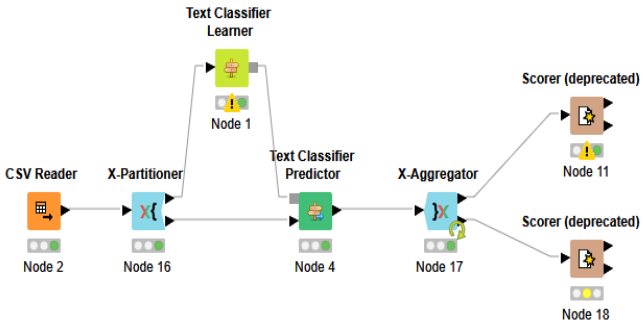

Figura 1. O *pipeline* do KNIME usado como modelo de aprendizagem e classificação de mensagens tóxicas.

III. Resultados

Os achados obtidos por este estudo preliminar de acordo com a metodologia apresentada na Seção II são expostos nas tabelas a seguir.

A Tabela I apresenta a matriz de confusão do experimento, onde o valor um (1) refere-se às mensagens consideradas tóxicas e o valor zero (0) às não-tóxicas, de acordo com a avaliação dos especialistas. À esquerda, os 0 e 1 representam os rótulos verdadeiros das mensagens e a primeira linha da tabela informam como o modelo as classificou. Assim, temos que o modelo alcançou uma acurácia média de 87% para esse contexto. Além disso, os resultados também são satisfatórios quando analisamos os erros, já que para o problema de classificação de mensagens tóxicas os falsos negativos são mais importantes que os falsos positivos, e somente 246 mensagens foram classificadas como não-tóxicas sendo que continham toxicidade.

Tabela I
Matriz de Confusão

		Previsto	
Real	0	3205	776
	1	247	4410

Já as métricas avançadas de classificação como Revocação, Precisão, Sensibilidade, Especificidade e Medida-F, são exibidas na Tabela II. Para medir a confiabilidade do experimento, o coeficiente Kappa foi utilizado e obteve um valor de 0,75, considerando o resultado como *substancial* [15].

Tabela II
MÉTRICAS AVANÇADAS

	Revocação	Precisão	Sensibilidade	Especificidade	Med.-F
0	0.805	0.925	0.805	0.945	0.861
1	0.946	0.846	0.947	0.801	0.894

IV. DISCUSSÃO

A precisão é definida como a fração de itens corretamente classificados dentro do total de itens recuperados em uma consulta [16], por exemplo, o percentual de mensagens realmente não-tóxicas recuperadas ao buscar por mensagens não-tóxicas na base de dados do estudo. Na área de busca e recuperação de informação, a revocação é o percentual dos documentos recuperados que constam no resultado de uma consulta [16]. Se um total de 100 documentos for relevante a uma busca e apenas 30 foram recuperados a revocação daquela consulta é de 0,30. Como visto nos resultados, tanto a revocação quanto a precisão foram altas para ambas as classes de toxicidade, sempre com uma taxa superior a 0,80. Um destaque entre esses resultados é a revocação de mensagens tóxicas, que recuperou 95% do total de tais mensagens. Outra medida usada para avaliar o desempenho na tarefa de busca de informação é a medida-f, que usa uma média ponderada harmônica das duas medidas previamente descritas, a precisão e a revocação, e um peso β, que determina quanto cada componente deve ter na medida [17]. O valor de β usado para esse estudo é um (1), dando um peso igual à revocação e precisão [17]. A medida-F é relativamente alta tanto para mensagens tóxicas quanto não tóxicas, com 0,84 e 0,89, respectivamente. O desempenho melhor na categoria de mensagens tóxicas talvez possa ser explicada pelo uso frequente de palavrões, expressões racistas e homofóbicas e termos tóxicos no contexto de jogos, como ”ez”, que sempre foram classificados como tóxicos na pré-classificação.

V. CONCLUSÃO E TRABALHOS FUTUROS

Uma possível melhoria na classificação realizada nesse estudo seria diferenciar uma mensagem neutra de uma positiva e avaliar quão tóxico é uma mensagem. Um modelo de aprendizado com mais nós na saída seria necessário para fugir da classificação binária da toxicidade e ter um ranking do nível de toxicidade de mensagens. Esse ranking de toxicidade poderia ter os seguintes níveis: positivo, vagamente positivo, neutra, vagamente tóxica, tóxica e mensagem de ódio. O aumento em granularidade provavelmente resultaria em uma acurácia mais baixa no modelo como um todo, porém traria informações sobre mensagens positivas e também identificaria discurso de ódio.

DOTA 2 é um jogo competitivo entre duas equipes. Por este motivo, ele possui dois canais de comunicação:

- Geral, dedicado à comunicação entre todos os jogadores da partida, para que eles troquem informações sobre a mesma.
- Equipe, dedicado à comunicação entre os jogadores de uma mesma equipe, para discussão das estratégias que serão adotadas pelo time.

Apesar dos objetivos associados a cada um dos canais, muitas vezes eles são desvirtuados. O geral se torna um meio para humilhar e insultar o time adversário, quando este está perdendo, por exemplo. E o de equipe passa a ser utilizado para ofender os outros membros.A toxicidade no chat de jogos acontece quando uma mensagem de conotação negativa e conteúdo pejorativo é dita na relação jogador-jogador, fugindo do caráter esportivo esperado de um jogo e inclui, mas não se estende a racismo, homofobia entre outros atos inaceitáveis em qualquer segmento da sociedade. É importante pontuar que existem mensagens tóxicas particulares a jogos, como “ez” que significa que a vitória sobre o oponente foi fácil.

A marca de 92% para mensagens não tóxicas e 85% para mensagens tóxicas na precisão com o uso de 8784 mensagens pré-classificadas para o treinamento mostra que é possível determinar a toxicidade de mensagens em jogos MOBA. Esse resultado positivo possibilita que algumas pesquisas futuras provenientes da classificação de toxicidade em chat de jogos possam ser feitas com base nos dados e modelos de aprendizado usados para esse estudo.Os atributos incluídos no arquivo usado para esse estudo ou de uma base de dados personalizada a partir de requisições à API do OpenDota abrem um leque de alternativas para correlacionar com a toxicidade das mensagem. As possibilidades geradas por uma nova base de dados personalizada através dos dados públicos do OpenDota são, mas não se limitam:

- Os nomes dos usuários permitem que sejam identificados os jogadores que emitem toxicidade durante o jogo;
- O horário relativo ao início da partida também pode motivar pesquisas futuras, pois poderíamos identificar uma mudança no comportamento dos jogadores antes do jogo comparado a durante o jogo;
- outra pesquisa relacionada ao horário identificaria como os jogadores reagem a resultados favoráveis ou desfavoráveis ao longo do jogo;
- Seguindo a linha dos resultados gerais do time, o OpenDota também disponibiliza dados sobre o desempenho dos jogadores, que permitem correlacionar mortes, dano, assistências e outras informações dos jogadores com a toxicidade;
- A possibilidade de se realizar um estudo sobre análise de sentimentos durante as trocas de mensagens nas partidas é uma possibilidade;
- A plataforma OpenDota fornece dados de partidas profissionais, ligas regionais e as chamadas partidas casuais, isso poderia permitir um estudo que explorasse o grau de toxicidade entre esses tipos de partida, indicando onde um jogador ficaria mais exporto a este tipo de ação.

O potencial demonstrado através deste estudo pode permitir a construção de ferramentas automatizadas e baseadas em inteligência que concretizem ações reais contra o comportamento tóxico em comunidades de jogos multijogador online.

Permitindo uma série de ações que possam cada vez mais potencializar o mercado e garantir um ambiente mais trativo, Corroborando com outros trabalhos desenvolvidos anteriormente neste sentido.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil - RESOLUÇÃO NORMATIVA RN-017/2006 e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

REFERÊNCIAS

[1] NewZoo, "Most popular core pc games — global," Tech. Rep., 2019. [Online]. Available: https://newzoo.com/insights/rankings/top-20-core-pc-games/

[2] M. Mora-Cantallops and M.-Á. Sicilia, "Moba games: A literature review," *Entertainment computing*, vol. 26, pp. 128–138, 2018.

[3] N. company, "2019 year in review: Digital games and interactive media," Tech. Rep., 2019. [Online]. Available: https://www.superdataresearch.com/reports/p/2019-year-in-review

[4] J. Blackburn and H. Kwak, "Stfu noob! predicting crowdsourced decisions on toxic behavior in online games," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 877–888.

[5] H. Kwak and J. Blackburn, "Linguistic analysis of toxic behavior in an online video game," in *International Conference on Social Informatics*. Springer, 2014, pp. 209–217.

[6] J. Y. Shim, T. H. Kim, and S. W. Kim, "Decision support of bad player identification in moba games using pagerank based evidence accumulation and normal distribution based confidence interval," *International Journal of Multimedia & Ubiquitous Engineering*, vol. 9, no. 8, pp. 13–16, 2014.

[7] M. Märtens, S. Shen, A. Iosup, and F. Kuipers, "Toxicity detection in multiplayer online games," in *Proceedings of the 2015 International Workshop on Network and Systems Support for Games*, ser. NetGames '15. IEEE Press, 2015.

[8] Opendota api (v18.0.0). [Online]. Available: https://docs.opendota.com/

[9] kaggle, comunidade on-line de cientistas de dados. [Online]. Available: https://www.kaggle.com/,

[10] Dota 2 matches dataset: Base de dados.

[11] K. partitioner. (2020) X-partitioner. [Online]. Available: https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.meta.xvalidation.XValidatePartitionerFactory

[12] K. T. Classifier. (2020) Text classifier. [Online]. Available: https://www.knime.com/book/text-classifier

[13] K. aggregator. (2020) X-aggregator. [Online]. Available: https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.meta.xvalidation.AggregateOutputNodeFactory

[14] W. Tang, J. Hu, H. Zhang, P. Wu, and H. He, "Kappa coefficient: a popular measure of rater agreement," *Shanghai archives of psychiatry*, vol. 27, no. 1, pp. 62–67, Feb 2015, 25852260[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25852260

[15] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012, 23092060[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23092060

[16] O. N. P. Cardoso, "Recuperação de informação," *INFOCOMP Journal of Computer Science*, vol. 2, no. 1.

[17] P. F. Matos, L. de Oliveira Lombardi, R. R. Ciferri, P. D. T. A. S. Pardo, C. D. de Aguiar Ciferri, and M. T. P. Vieira, "Relatório técnico "métricas de avaliação"," Tech. Rep., November 2009.