CLASSIFICAÇÃO DE VALORES HUMANOS BASEADO EM TEXTO USANDO APRENDIZAGEM DE MÁQUINA[1]

Sérgio Ewerton Barbosa Correia, Yuri de Almeida Malheiros Barbosa

Departamento de Ciências Exatas - Universidade Federal da Paraíba (UFPB)
Rio Tinto - PB - Brasil

{sergio.ewerton, yuri}@dce.ufpb.br

Abstract. *Twitter receives millions of messages daily from different users and regions of the world. These messages contain opinions, evaluations and feelings, making Twitter an ideal platform for various types of study. Analyzing shared texts on social networks can show us various kinds of information about a person, emotional state, opinions about products, political preference. This work aims to create a model for classification of human values using machine learning to predict the results according to textual messages collected from Twitter. Four classifiers were evaluated, reaching a maximum value of 95,16% of accuracy.*

Resumo. O Twitter recebe diariamente milhões de mensagens de diferentes usuários e regiões do mundo. Mensagens estas contendo opiniões, avaliações e sentimentos, tornando o Twitter uma plataforma ideal para diversos tipos de estudo. Analisar textos compartilhados em redes sociais pode nos mostrar diversos tipos de informações de uma pessoa, estado emocional, opiniões sobre produtos, preferência política. Este trabalho tem como objetivo criar um modelo para classificação de valores humanos usando aprendizagem de máquina para prever os resultados de acordo com mensagens textuais coletadas do Twitter. Foram avaliados quatro classificadores atingindo um valor máximo de 95,16% de acerto.

1. Introdução

Cada vez mais pessoas utilizam redes sociais para expor opiniões e sentimentos, sendo o Twitter uma das plataformas mais utilizadas para esse fim. Com posts de até 280 caracteres e cerca de 335 milhões de usuários mensalmente ativos, o Twitter recebe uma enorme quantidade de conteúdo diariamente se tornando um ambiente perfeito para diversos tipos de pesquisas (Twitter, 2018). Além disso, o Twitter disponibiliza uma plataforma para desenvolvedores no qual oferece diversas APIs, ferramentas e recursos que possibilitam aproveitar essa rede social de uma forma automatizada.

Analisar textos compartilhados em redes sociais pode nos mostrar diversos tipos de informações de uma pessoa, estado emocional, opiniões sobre produtos, preferência

[1] Trabalho de Conclusão de Curso (TCC) na modalidade Artigo apresentado como parte dos pré-requisitos para a obtenção do título de Bacharel em Sistemas de Informação pelo curso de Bacharelado em Sistemas de Informação do Centro de Ciências Aplicadas e Educação (CCAE), Campus IV da Universidade Federal da Paraíba, sob a orientação do professor Yuri de Almeida Malheiros Barbosa.

política e, até mesmo, encontrar sinais de distúrbios mentais. Sabendo disso, a psicologia tem um papel fundamental no entendimento desses fenômenos. Segundo (Gouveia, et al. 2009, p. 36), "Os valores humanos são um construto central na psicologia (Rokeach, 1973) e têm relevância especial para o entendimento de diversos fenômenos sociopsicológicos".

Dentro da Psicologia Social existe a teoria dos valores humanos, na qual atualmente consegue descobrir os valores de uma pessoa através da resposta de um questionário com 18 itens (Medeiros, 2011). Esta teoria identifica duas funções consensuais dos valores. A primeira com três tipos de orientações valorativa: social, central e pessoal, e na segunda com dois tipos: humanitário e materialista. Como pode ser visto na Figura 1, a junção dessas duas funções cria seis subfunções de valores (normativa, realização, existência, suprapessoal, interacional e experimentação) (Gouveia, et al. 2009).

Este trabalho tem como proposta criar um modelo de classificação para valores humanos para analisar a viabilidade de usar aprendizagem de máquina para prever os resultados de acordo com mensagens textuais coletadas do Twitter. Assim, sendo possível definir os valores humanos de uma pessoa apenas com sua autorização para coleta de suas mensagens, sem que seja necessário a adição de uma ou mais atividades extras.

O trabalho estrutura-se da seguinte forma: Na segunda seção, são apresentados conceitos essenciais para a compreensão do trabalho, por exemplo, valores humanos e aprendizagem de máquina. Na terceira seção, apresenta-se a metodologia aplicada, descrevendo o processo da coleta dos dados, pré-processamento dos dados e treinamento do classificador. Na quarta, são apresentados os resultados alcançados. Na quinta, discussões. Na sexta e última seção, são apresentados a conclusão e os trabalhos futuros.

2. Fundamentação teórica

2.1 Valores humanos

Valores humanos são de grande importância por servir de construtor para a compreensão de inúmeros fenômenos sócio-psicológicos. Segundo (Medeiros, 2011), essa teoria tem sido empregada para esclarecer atitudes e comportamentos ambientais, religiosidade, preconceito, consumo de drogas, comportamento antissociais, delinquência juvenil, atitudes frente à tatuagem, intenção de cometer suicídio, uso de água e atitudes pró-ambientais.

Os valores humanos em Psicologia Social são definidos como princípios guias gerais e segundo (Medeiros, 2011, p. 75) "[...] servem como (1) critérios de orientação que guiam o comportamento das pessoas e (2) expressam cognitivamente suas necessidades". As pesquisas sobre valores humanos ou trabalhos que aplicam essa teoria têm ganhado lugar de destaque na Psicologia Social, provavelmente por conseguir realizar uma importante função no processo seletivo das atitudes humanas (Medeiros, 2011). Como exemplo, no trabalho de (Gouveia, et al. 2009), foram mostrados os

resultados de três estudos práticos, demonstrando a aplicabilidade da teoria funcionalista dos valores para a gestão de pessoas e de organizações. Os estudos indicaram que existe relação das subfunções dos valores com comprometimento organizacional, bem-estar afetivo no trabalho, *burnout* e fadiga.

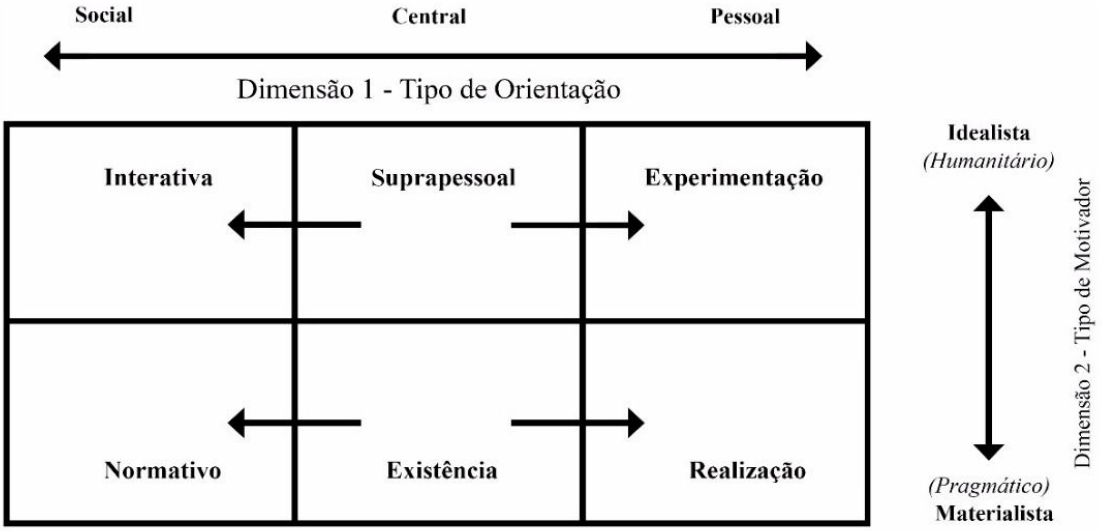

Figura 1. Matriz das duas funções consensuais dos valores
Fonte: (Gouveia, et al. 2009)

O questionário dos Valores Básicos foi inicialmente desenvolvido em português e espanhol contendo 66 itens. Em seguida, foram criadas duas versões abreviadas em português. A primeira deixou o questionário com apenas com 24 itens e na segunda esse número caiu para 18, sendo esta a mais utilizada atualmente (Medeiros, 2011). Cada um dos 18 itens do questionário pode ser respondido utilizando uma escala que varia de 1 (Totalmente não importante) a 7 (Totalmente importante).

A teoria funcionalista dos valores humanos também conta com uma estrutura definida para as funções e subfunções dos valores, como podemos ver na Figura 1. Essa estrutura é a combinação de duas dimensões principais, a primeira no eixo horizontal, correspondendo ao tipo de orientação (social, central ou pessoal) e a segunda representada no eixo vertical, correspondendo aos tipos de motivadores (materialista ou humanitário). As setas que saem do centro para as laterais representam a referência da fonte principal das subfunções (Gouveia, et al. 2009). A junção dessas duas funções gera seis subfunções de valores, que são os possíveis resultados para o questionário e estão detalhados na Tabela 1.

SUBFUNÇÕES VALORATIVAS	MOTIVADORES E ORIENTAÇÕES	VALORES BÁSICOS E SUAS DESCRIÇÕES
Experimentação	Motivador humanitário e orientação pessoal	EMOÇÃO. Desfrutar desafiando o perigo; buscar aventuras. PRAZER. Desfrutar a vida; satisfazer todos os seus desejos. SEXUALIDADE. Ter relações sexuais; obter prazer sexual.

Realização	Motivador materialístico e orientação pessoal	PODER. Ter poder para influenciar os outros e controlar decisões; ser o chefe de uma equipe. PRESTÍGIO. Saber que muita gente o conhece e admira; quando velho, receber uma homenagem por suas contribuições. ÊXITO. Obter o que se propõe; ser eficiente em tudo que faz.
Existência	Motivador materialístico e orientação central	SAÚDE. Preocupar-se com sua saúde antes mesmo de ficar doente; não estar enfermo. ESTABILIDADE PESSOAL. Ter certeza de que amanhã terá tudo o que tem hoje; ter uma vida organizada e planificada. SOBREVIVÊNCIA. Ter água e comida, e poder dormir bem todos os dias; viver em um lugar com abundância de alimentos.
Suprapessoal	Motivador humanitário e orientação central	BELEZA. Ser capaz de apreciar o melhor da arte, música e literatura; ir a museus ou exposições onde possa ver coisas belas. CONHECIMENTO. Procurar notícias atualizadas sobre assuntos pouco conhecidos; tentar descobrir coisas novas sobre o mundo. MATURIDADE. Sentir que conseguiu alcançar seus objetivos na vida; desenvolver todas as suas capacidades.
Interativa	Motivador humanitário e orientação social	AFETIVIDADE. Ter uma relação de afeto profunda e duradoura; ter alguém para compartilhar seus êxitos e fracassos. CONVIVÊNCIA. Conviver diariamente com os vizinhos; fazer parte de algum grupo, como: social, religioso, esportivo, entre outros. APOIO SOCIAL. Obter ajuda quando necessitar; sentir que não está só no mundo.
Normativa	Motivador materialístico e orientação social	OBEDIÊNCIA. Cumprir seus deveres e obrigações do dia a dia; respeitar seus pais, os superiores e os mais velhos. RELIGIOSIDADE. Crer em Deus como o salvador da humanidade; cumprir a vontade de Deus. TRADIÇÃO. Seguir as normas sociais de seu país; respeitar as tradições de sua sociedade.

Tabela 1 - Explicação das 6 subfunções valorativas.
Fonte: (Gouveia, et al. 2009)

2.2 Aprendizagem de máquina (AM)

Aprendizagem de Máquina é uma subárea de inteligência artificial, na qual um programa, através da experiência (novos dados) consegue melhorar seu desempenho em diversas funções, assim tomando decisões melhores de forma automatizada e sem a interferência humana (Santos, 2016). Por exemplo, um jogo de xadrez que utiliza um algoritmo de AM a cada partida ele ganha novos dados sobre jogadas e estratégias, obtendo assim experiências e se aperfeiçoando a cada jogo. Os três principais tipos de aprendizagem de máquina são: aprendizagem não supervisionada, aprendizagem supervisionada e aprendizagem por reforço (Norvig, Russell, 2014). A seguir uma breve descrição de cada um deles.

Na Aprendizagem não supervisionada, os dados de treinamento não recebem rótulos ou *feedback*, o algoritmo aprende padrões de entrada e realiza o agrupamento dos dados, sendo essa a função mais comum na aprendizagem não supervisionada (Norvig, Russell, 2014). Posteriormente se faz necessário uma análise para estabelecer o que cada agrupamento representa no contexto da situação onde foi implementado [Santos, 2016].

Na aprendizagem supervisionada, para cada exemplo, existe um rótulo ou saída relacionado. Os exemplos são mencionados como supervisionados, pois, além de cada exemplo conter suas características, contém também um rótulo de saída (Santos, 2016).Assim, um modelo pode ser treinado baseado nos dados rotulados para classificar novas entradas. Esse tipo de abordagem é muito utilizado para problemas de classificação como a detecção de spam, por exemplo. Quando um usuário marca um email como spam ele está dando um rótulo para aquele email. Com isso, o algoritmo monta uma base para tentar identificar novas entradas. Logo, quanto maior for essa base mais preciso o algoritmo tende a ser para classificar novas entradas não rotuladas.

A metodologia da aprendizagem por reforço utiliza um sistema de recompensas ou punições para se aperfeiçoar no ambiente que se encontra (Norvig, Russell, 2014). Por exemplo, a falta de gorjeta ao final de um atendimento indica ao garçom que algo saiu errado. Os dois pontos no final de um jogo de xadrez indica que o jogador tomou decisões corretas. Sabendo disso, compete ao agente decidir qual das ações realizadas foram responsáveis para a recompensa ou punição (Norvig, Russell, 2014).

Além da escolha do tipo de aprendizagem, uma das etapas mais importantes na aprendizagem de máquina é o pré-processamento dos dados, principalmente quando se diz respeito ao tratamento de texto (Benevenuto et al., 2015). Com essa etapa, conseguimos eliminar algumas características que não agregam muita informação. E, para isso, podemos utilizar algumas das técnicas descritas a seguir.

Stopwords é uma lista de palavras conhecidas que não agregam valor ao texto, normalmente fazem parte desta lista as preposições, pronomes, artigos, advérbios, e outras classes de palavras auxiliares (Morais e Ambrósio, 2007). Na língua portuguesa temos como exemplos de stopwords palavras como "já", "para" e "dos". Outro processo é o stemming, que tem como objetivo eliminar as variações morfológica das palavras e faz isso retirando prefixos, sufixos, características de gênero, número e grau, deixando apenas seu radical (Morais e Ambrósio, 2007). Temos como exemplo as palavras "observar", "observadores", "observasse", "observou" e "observe" que podem ser resumidas para seu radical "observ".

Existem também dois problemas comuns que podem ocorrer durante o treinamento de um modelo. O Overfitting que, segundo (Martins, 2003), ocorre quando o classificador "decora" os dados de treinamento, apresentando um alto grau de precisão no treinamento e uma alta taxa de erros para os casos de teste. E, por outro lado temos o Underfitting, que para (Martins, 2003) normalmente acontece quando a base de dados para o treinamento é muito pequena, fazendo com que o classificador não consiga

abstrair o conceito, assim apresentando um alto grau de erros tanto do treinamento quando nos casos de teste.

3. Metodologia

O presente trabalho tem como objetivo criar um classificador para valores humanos baseado em mensagens textuais compartilhadas no Twitter e, para isso, foram realizadas as seguintes etapas.

3.1 Coleta dos dados

O Twitter disponibiliza uma plataforma para desenvolvedores que fornece diversas APIs, ferramentas e recursos que possibilitam aproveitar as funcionalidades da rede de forma automatizada. Entre as funções disponibilizadas por essas APIs, existe uma que retorna uma coleção dos tweets mais recentes postados por um determinado usuário. Sabendo disso, foi desenvolvida uma aplicação web utilizando a linguagem de programação Python (3.6) e o framework para desenvolvimento web Django (1.11) na qual o usuário pode realizar o login através de sua conta do Twitter e conceder autorização para que a aplicação possa realizar a captura de até 3200 tweets da sua timeline, sendo esse número a quantidade máxima estabelecida pela API para cada usuário. Essa quantidade é adequada para o propósito deste trabalho. Após a autorização, o usuário é redirecionado para um questionário com as 18 perguntas sobre valores humanos e de forma assíncrona é realizada a captura dos tweets. Ao finalizar o questionário a aplicação persiste as respostas e os tweets coletados em um banco de dados. A aplicação web está disponível em https://valoreshumanos.dcx.ufpb.br, na Figura 2 podemos ver sua página inicial e a tela da primeira pergunta do questionário.

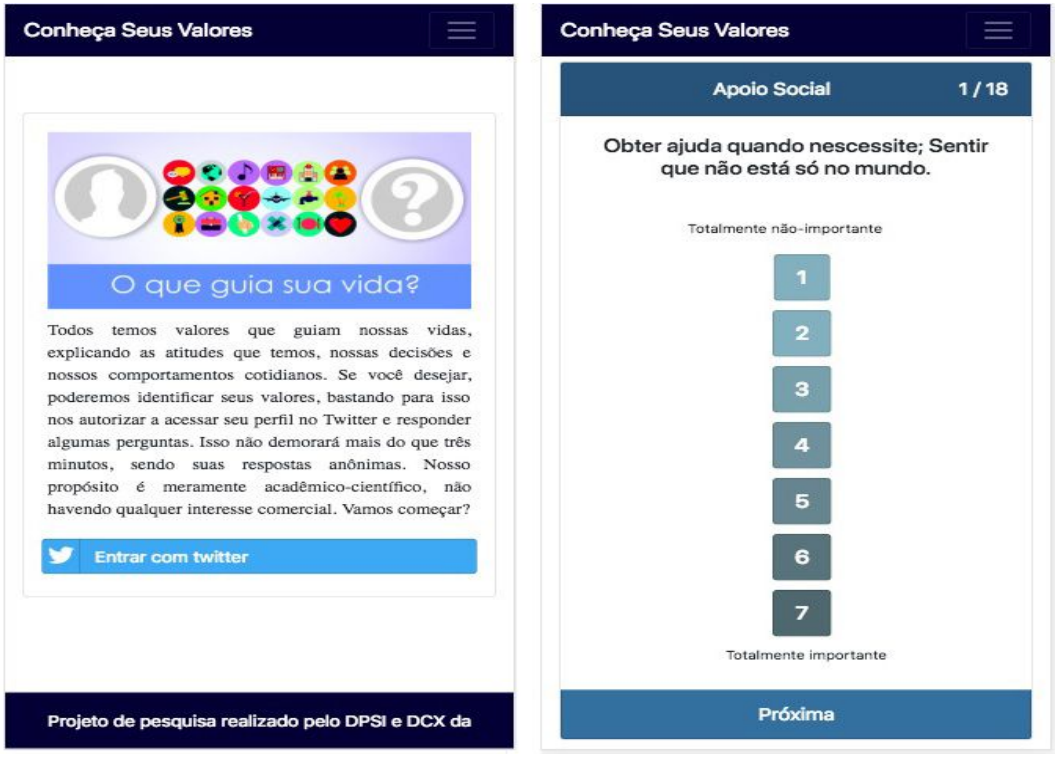

Figura 2. Página inicial e primeira pergunta do questionário

Como podemos observar na Tabela 2, foram coletados dados de 2.496 usuários, entretanto foi necessário excluir os usuários que não responderam o questionário completamente, uma vez que só é possível identificar seus valores humanos com o questionário totalmente respondido, e também os usuários que não possuíam tweets em sua timeline. Posteriormente na Tabela 3, podemos observar a quantidade de usuários e tweets por classe. Podemos observar também, que existe um desbalanceamento em relação a quantidade de tweets, por exemplo, a classe normativa com 88.090 e suprapessoal com 596.239 tweets.

Tabela 2. Quantidade de usuários e quantidade de tweets após cada filtro

	Usuário	Tweets
Total de usuários que realizaram o login no sistema	2.496	2.277.832
Que responderam o questionário completo	2.048	1.934.971
Que tinha pelos menos 1 tweet postado	1.975	1.934.971

Tabela 3. Quantidade de usuários e tweets por classe

Nome	Quantidade Usuários	Quantidade Tweets
Existência	520	500.157
Experimentação	275	276.489
Interativa	170	176.569
Normativa	107	88.090
Realização	281	297.427
Suprapessoal	622	596.239
Total	1.975	1.934.971

3.2 Pré-processamento dos dados

Essa etapa tem como objetivo reduzir o número de palavras, remover caracteres indesejados, tais como sinais de pontuação, marcadores, números, entre outros, os quais separadamente não agregam valor para o algoritmo de aprendizagem. Antes de aplicar qualquer tratamento nos textos foi realizada a tokenização, que é a quebra do texto em palavras ou tokens. Em seguida, foram realizadas as seguintes etapa de pré-processamento em cada *token*. Foi removido caracteres especiais (@, #, ") e números através de expressões regulares, depois disso, foi utilizada a biblioteca NLTK que contém a lista dos *stopwords* em português, com essa lista foi possível remover as palavras que não agregam valor ao documento. E, por último, e ainda utilizando a mesma biblioteca, foi realizado o processo de *stemming*, que elimina as variações morfológica deixamos apenas o radical de cada palavra.

A seguir tem-se dois exemplos de tweets antes e depois do pré-processamento.

Entrada 1: '"Minha expectativa pra receber notas de provas nunca vão ser positivas, tipo nunca"'
Saída 1: 'expect pra receb not prov nunc va ser posit tip nunc '

Entrada 2: '"RT @paopixel: Tô numa quadrilha e gritaram "Olha o Neymar" e todos os homens deitaram no chão \n\n O MELHOR DO BRASIL é O BRASILEIRO"'
Saída 2: 'rt paopixel to quadrilh grit olha neym tod homens deit cho melhor brasil brasileir '

3.3 Treinamento

Durante o treinamento, foi identificado um desbalanceamento em relação a quantidade de tweets para cada classe. Segundo (Santos, 2016), quando essa diferença é grande, os algoritmos de aprendizado tornam-se tendenciosos para um determinado resultado, prejudicando assim o resultado final. Sabendo disso, foi utilizada a técnica *under-sampling* que é uma solução direta para o problema de classes desbalanceadas (Batista, 2005). Essa técnica consiste na identificação da menor classe e a exclusão de forma aleatória de uma parcela dos dados das classes majoritárias, com a intenção de realizar o balanceamento tomando como referência a menor. Ao finalizar o balanceamento, foram utilizados quatro algoritmos de aprendizagem de máquina supervisionada: Logistic Regression, Naive Bayes, Random Forest e Linear SVC, todos implementados pela biblioteca Scikit-Learn.

Sabendo que cada tweet tem no máximo 280 caracteres e que após passar pelo pré-processamento ele apresenta uma perda significativa de seu tamanho, foi utilizado a estratégia de agrupar tweets de uma mesma classe de resultado, com o objetivo de investigar como esses agrupamentos alteram o resultado dos classificadores. Foram utilizadas as seguintes combinações: 10, 20, 40, 60, 80, 100, 120, 150 e 180. Assim, cada exemplo do treinamento tem o conteúdo de X tweets e o rótulo da resposta do questionário. Todos os resultados estão expostos na Tabela 4.

Após o treinamento do classificador, podem ocorrer dois problemas conhecidos na aprendizagem de máquina, o Overfitting que ocorre quando o classificador apresenta um alto grau de precisão no treinamento e uma alta taxa de erro para os casos de teste. E também o Underfitting, que acontece quando o algoritmo apresenta um alto grau de erro tanto no treinamento quando nos casos de teste. Sabendo disso e com o objetivo de validar o classificador, foi utilizada a estratégia de Cross-Validation ou validação cruzada, que consiste na divisão do dataset em duas partes, treinamento (90% dos dados) e teste (10% dos dados). Essa divisão é realizada de forma aleatória e esse procedimento foi repetido 5 vezes, e a cada ciclo foi realizada a execução dos 4 classificadores, com a finalidade de analisar o resultado individual fornecendo a eles os mesmos dados.

4. Resultados

Na Tabela 4, podemos observar as porcentagens de acerto de cada algoritmo, sabendo que cada porcentagem é a média de 5 execuções utilizando a validação cruzada. Na primeira linha observamos a quantidade de tweets combinados em cada exemplo de treinamento e na última linha a quantidade de exemplos restante após essa combinação. O resultado de cada algoritmo em relação ao número de tweets agrupados também pode ser observado de forma gráfica na Figura 3. Obtivemos resultados positivos com a união de alguns tweets e a partir de 40 tweets agrupados, o aumento na porcentagem de acertos não foram muito significativos, além de que a quantidade de tweets para cada classe de resultado diminui.

Tabela 4. Porcentagens de acertos em relação ao agrupamento dos tweets.

Quantidade de tweets por agrupamento	**1**	**10**	**20**	**40**	**60**	**80**	**100**	**120**	**150**	**180**
Logistic Regression	59,38%	80,66%	86,56%	90,53%	90,92%	92,70%	93,45%	94,29%	94,26%	95,16%
Naive Bayes	59,50%	76,20%	71,02%	81,90%	82,06%	79,61%	77,20%	76,39%	78,47%	78,02%
RandomForest Classifier	52,60%	68,60%	75,22%	74,20%	74,97%	74,27%	74,77%	72,20%	71,48%	70,03%
Linear SVC	61,78%	82,70%	88,17%	92,32%	93,61%	94,37%	94,02%	94,71%	94,77%	94,95%
Quantidade de exemplos em cada classe	88090	8809	4405	2202	1468	1101	881	734	587	489

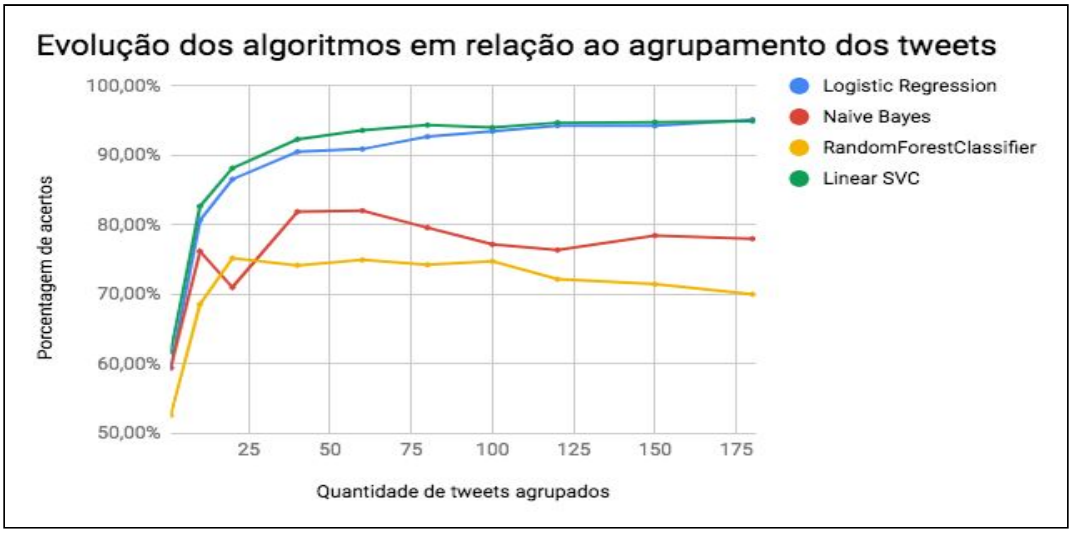

Figura 3. Representação gráfica da Tabela 4

5. Discussão

Como podemos observar na Tabela 4, obtivemos diversas porcentagens de acertos para quatro algoritmos distintos e para nove quantidades diferentes de tweets agrupados. Os testes foram feitos com o intuito de analisarmos qual seria a melhor quantidade para o agrupamento e qual algoritmo teria o melhor desempenho.

Analisando a Figura 3 vemos que a partir de 80 tweets combinados é muito pequeno o aumento da porcentagem de acertos nos classificadores Logistic Regression e Linear SVC, já nos algoritmos Random Forest e Naive Bayes os resultados pioram. Com isso, temos a indicação de que um agrupamento em torno de 80 tweets é adequado para o propósito do trabalho. Além disso, levando em consideração que a quantidade de tweets combinados é inversamente proporcional a quantidade de exemplos no dataset, aumentando muito o tamanho dos agrupamentos, podemos causar o problema de Underfitting.

Poderíamos ter resultados diferentes com o agrupamento de mais tweets, mas não conseguimos aumentar essa quantidade devido ao desbalanceamento das classes, que deixou o número total de exemplos reduzido. Como por exemplo, a classe "Normativa" teve um total de 88.090 tweets, 676,8% a menos que a classe "Suprapessoal" com 528.540. E após realizar o balanceamento, todas as outras classe igualaram essa quantidade, deixando o dataset com 528.540 (88.090 * 6) exemplos.

Observamos também, que a porcentagem de acertos nos classificadores Logistic Regression e Linear SVC foram as melhores e se mantiveram bem próximas. Em uma possível escolha entre eles, poderíamos utilizar para o desempate o tempo de execução, no qual o algoritmo Linear SVC tem um tempo de execução em uma máquina de configuração (sistema operacional mac OS x, processador i5 e memória de 8gb) três vezes menor que o Logistic Regression com 44 segundos.

Como podemos observar nos dados, o classificador conseguiu encontrar relações entre os o que os usuários escrevem e os seus valores humanos retornados pelo questionário. Um acerto de mais de 90% para 6 classes, mostra que com certeza foram encontrados padrões nos textos. Com 6 classes, selecionar aleatoriamente um dos resultados teria 16,666% de chance de acerto, que não foi o caso. Tivemos algoritmos que chegaram acima de 90% e essa alta taxa de acerto é um indício forte que podemos substituir o questionário por uma análise textual.

6. Conclusão e trabalhos futuros

Este trabalho apresentou a criação de um classificador para valores humanos, com a finalidade de analisar a viabilidade de usar aprendizagem de máquina para os valores humanos das pessoas de acordo com mensagens textuais coletadas do Twitter. Foram descritos os processos para coleta dos dados, em que conseguimos coletar um total de 2.496 usuários e 2.277.832 tweets. Pré-processamento dos dados, que contou com a remoção dos stopwords, aplicação do processo de stemming e remoção de caracteres especiais que não agregam valor ao classificador. O balanceamento das classes de resultado, com o objetivo de manter a igualdade durante o treinamento. Também

utilizamos a estratégia de agrupar tweets de uma mesma classe de resultado, com o objetivo de analisar o comportamento do classificador. E por fim foi realizado o treinamento, no qual contou com a execução de quatro algoritmos de aprendizagem de máquina supervisionado: Logistic Regression, Naive Bayes, Random Forest e Linear SVC, todos implementados pela biblioteca Scikit-Learn.

Conseguimos excelentes resultados com todos os algoritmos, mas em especial com o Linear SVC, que obteve uma média de 94,77% de acertos, ou seja, dos 110 casos de teste o algoritmo classificou corretamente 103. Com esse resultado podemos confirmar que existe viabilidade no uso de modelos textuais criados por algoritmos de aprendizagem de máquina como instrumento para identificar os valores humanos das pessoas. Podendo otimizar o tempo dos usuários e, até mesmo, aumentar o número de participantes válidos. Tendo em vista que durante o pré-processamento dos dados, um dos filtros aplicados foi para remoção dos usuários que não responderam completamente o questionário, fez a exclusão de 448 (17,9%) dos usuários, ou seja, em um questionário de três a quatro minutos aproximadamente de duração, tivemos essa porcentagem considerável de abandono, problema que não aconteceria com o uso exclusivo do classificador.

Como trabalhos futuros, pretende-se repetir o processo de coleta de dados, com o objetivo de aumentar a quantidade de tweets da classe normativa, por ser ela no processo de balanceamento a responsável pela exclusão de uma grande parcela dos dados das outras classes. A execução de outros algoritmos de aprendizagem de máquina para analisar seu desempenho com o dataset criado. Também pretende-se replicar a mesma ideia deste trabalho para a rede social Facebook, tendo em vista um número ainda maior de usuário mensalmentes ativos, aproximadamente 2,23 bilhões de usuário (Statista, 2018). E, por fim, a implementação de uma API para uso do classificador por desenvolvedores em diversos projetos.

REFERÊNCIAS

BATISTA, Gustavo Enrique de Almeida Prado. Pré-processamento de dados em aprendizado de máquina supervisionado. 2003. Tese de Doutorado. Universidade de São Paulo.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para análise de sentimentos em mídias sociais. 2015.

BITTENCOURT, Valnaide Gomes. Aplicação de técnicas de aprendizado de máquina no reconhecimento de classes estruturais de proteínas. 2005. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte.

GOUVEIA, Valdiney V. Teoria funcionalista dos valores humanos: Aplicações para organizações. Revista de Administração Mackenzie (Mackenzie Management Review), v. 10, n. 3, 2009.

MARTINS, Cláudia Aparecida. Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado. 2003. Tese de Doutorado. Universidade de São Paulo.

MEDEIROS, Emerson Diógenes de et al. Teoria funcionalista dos valores humanos: Testando sua adequação intra e interculturalmente. 2011.

MICHALSKI, Ryszard S.; CARBONELL, Jaime G.; MITCHELL, Tom M. (Ed.). Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. Mineração de textos. Relatório Técnico–Instituto de Informática (UFG), 2007.

NORVIG, Peter; RUSSELL, Stuart. Inteligência Artificial: Tradução da 3a Edição. Elsevier Brasil, 2014.

ROKEACH, Milton. The nature of human values. Free press, 1973.

SANTOS, Rodrigo Magalhães Mota dos. Técnicas de aprendizagem de máquina utilizadas na previsão de desempenho acadêmico. 2016. Dissertação de Mestrado. UFVJM.

Statista, (2018). Number of monthly active Facebook users worldwide as of 2nd quarter 2018 (in millions). Acesso em 17 de 10 de 2018, disponível em: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Twitter, (2018). Twitterinc investor. Acesso em 21 de 08 de 2018, disponível em : https://investor.twitterinc.com/financial-information