UNIVERSIDADE FEDERAL DE PERNAMBUCO

VALDEMIR DE ANDRADE BORGES JUNIOR

EXTRAÇÃO DE DADOS ATÍPICOS DE DISCURSOS DE PACIENTES COM DOENÇAS MENTAIS

Recife

2018

VALDEMIR DE ANDRADE BORGES JUNIOR

EXTRAÇÃO DE DADOS ATÍPICOS DE DISCURSOS DE PACIENTES COM DOENÇAS MENTAIS

Trabalho de Conclusão de Curso apresentado ao curso de Ciências da Computação, da Universidade Federal de Pernambuco, como requisito parcial para a Obtenção do grau de Bacharel em Ciências da Computação.

Recife

2018

VALDEMIR DE ANDRADE BORGES JUNIOR

EXTRAÇÃO DE DADOS ATÍPICOS DE DISCURSOS DE PACIENTES COM DOENÇAS MENTAIS

Trabalho de Conclusão de Curso apresentado ao curso de Ciências da Computação, da Universidade Federal de Pernambuco, como requisito parcial para a Obtenção do grau de Bacharel em Ciências da Computação.

Recife, 6 de Julho de 2018

BANCA EXAMINADORA

Katia Guimarães
Universidade Federal de Pernambuco

AGRADECIMENTOS

Agradeço primeiramente a minha mãe, por estar sempre presente e me ajudar no que podia, seja com materiais para o projeto ou simplesmente me ajudando a ter mais tempo livre. Sua paciência foi fundamental para a construção da minha carreira como estudante e para o meu progresso particularmente nesses últimos 6 meses.

Agradeço a professora Katia, minha orientadora nesse projeto, que além de ser uma excelente orientadora e me incentivar a buscar sempre melhorar, foi uma das melhores professoras da qual eu tive a felicidade de ser aluno em mais de uma ocasião.

Agradeço a Gabriel, presidente da Liga de Neurociências Aplicadas (LIANA), que, com os seus conhecimentos de neurofisiologia, foi de grande ajuda na parte mais biológica desse trabalho, onde eu, por ser aluno de ciências da computação e viver nesse mundo "dos computadores", teria dificuldades em encontrar conteúdo de confiança não fosse a sua prestativa e admirável cooperação.

Por fim, agradeço ao meu pequeno cachorro, Sheldon, que sempre soube me alegrar mesmo nos momentos mais complicados da elaboração desse trabalho e que me ensinou a apreciar os pequenos detalhes da vida.

RESUMO

O diagnóstico de doenças mentais em pacientes é um procedimento que requer precisão e rapidez. Tal procedimento é muito importante para essa área, pois além do fato de doenças mentais possuírem sintomas muito semelhantes, o tratamento delas difere muito entre si, e em muitos casos, o tratamento de uma doença mental não surte nenhum efeito no tratamento de outra. Usando técnicas de teoria dos grafos e de análise de sentimentos, foi desenvolvida uma técnica de análise de discurso que busca classificar um paciente baseado nos sinais de doenças mentais que o seu discurso apresenta.

Palavras-chave: Esquizofrenia. Teoria dos grafos. Inteligência artificial. Análise de sentimentos.

ABSTRACT

Mental health diagnosis in patients is a procedure that requires efficiency and quickness. Such procedure is very important in this area, because not only mental diseases have very similar symptoms, but they also have very different treatments, such that what works for one disease may have no effect for others. Using sentiment analysis and graph's theory techniques, a speech analysis technique was developed with the purpose of classifying a patient based on the signs of mental illnesses their speech presents.

Keywords: Schizophrenia, Graph theory, Artificial intelligence, Sentiment analysis.

LISTA DE ILUSTRAÇÕES

FIGURA 1 - ORGANIZAÇÃO EM CAMADAS 18
FIGURA 2 - OVERFITTING 19
FIGURA 3 - REPRESENTAÇÃO DE MODELOS SVC 22
FIGURA 4 - EXEMPLO DE TABELAS USADAS PELO MODELO GAUSSIANNB 24
FIGURA 5 - TRECHO DE CORPUS LÉXICO DA USP 27
FIGURA 6 - GRAFO COM PESOS 30
FIGURA 7 - GRAFO DIRECIONADO 30
GRÁFICO 8 - ACURÁCIA 40
GRÁFICO 9 – COMPARAÇÃO DOS RESULTADOS ENTRE OS GRUPOS 41
GRÁFICO 10 - COMPARAÇÃO DE CICLOS 42
GRÁFICO 11 - DADOS COMPLEMENTARES DOS GRAFOS 43

Sumário

1. INTRODUÇÃO 9

2. CONTEXTO 12

2.1 BASE DE DADOS 12

2.2 PROCESSAMENTO DE DADOS 13

3. LITERATURA ATUAL 15

4. INTELIGÊNCIA ARTIFICIAL 17

4.1 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL USADAS 20

5. ANÁLISE DE SENTIMENTOS 25

5.1 METODOLOGIA EM ANÁLISE DE SENTIMENTOS 28

6. TEORIA DOS GRAFOS 29

6.1 ALGORITMOS EM TEORIA DOS GRAFOS 32

6.2 TÉCNICAS DE TEORIA DOS GRAFOS USADAS 33

7. DOENÇAS MENTAIS ABORDADAS 34

7.1 ESQUIZOFRENIA 34

8. RELATÓRIO 37

8.1 ENTRADA/SAÍDA 38

8.2 FUNCIONAMENTO DO PROGRAMA 38

8.3 TESTES 39

TRABALHOS FUTUROS 45

CONCLUSÃO 46

APÊNDICE A — PERGUNTAS USADAS NAS ENTREVISTAS DE PACIENTES COM ESQUIZOFRENIA 47

ANEXO A — ALGORITMO DE JOHNSON 48

REFERÊNCIAS 52

1. INTRODUÇÃO

Um dos aspectos mais importantes e controversos da medicina atualmente é o diagnóstico de pacientes. Um diagnóstico correto é parte essencial do processo de tratamento e recuperação do indivíduo, ao mesmo tempo que um diagnóstico errado pode ser catastrófico ou atrasar o tratamento correto para um paciente em estado grave. Atualmente, mais de 12 milhões de pacientes são diagnosticados erroneamente todo ano só nos estados unidos, o equivalente a 5% de todos os pacientes diagnosticados (ABRIL,2016).

Uma das áreas mais afetadas por esta mazela é a área de psiquiatria, especialmente quando tratamos de diagnósticos prematuros. É extremamente difícil para um médico diagnosticar doenças mentais, especialmente quando consideramos que o escopo de personalidades humanas é tão elevado, o que torna praticamente impossível perceber o que constitui de fato um comportamento atípico, dentro de vícios e idiossincrasias adquiridas pelo paciente ao longo de sua vida. Um diagnóstico conclusivo às vezes só pode ser chegado uma vez que a doença já causou alterações na neuroanatomia grandes o suficiente para serem notadas através de neuroimagem, e a este ponto já pode ser tarde demais e essas alterações já podem ser irreversíveis.

Outro problema agravante nesse cenário é a análise de desenvolvimento do quadro clínico do paciente. Existem diversos fatores a serem analisados no comportamento do paciente, e alguns desses aspectos são subjetivos (o paciente relata se sentir melhor, embora sinta não estar totalmente curado). Em um cenário onde doenças com sintomas parecidos possuem tratamentos diferentes e ineficientes quando o diagnóstico está errado, e ainda levando em conta o efeito placebo, qualquer resultado que não seja 100% positivo pode trazer à tona a dúvida de que os sintomas que indicaram que o paciente sofria de autismo, por exemplo, talvez esteja sofrendo de demência, o que mudaria completamente o tratamento recomendado.

O diagnóstico prematuro de patologias mentais se torna essencial para uma recuperação de sucesso, e para isto, uma das ferramentas mais úteis para psiquiatras é a entrevista com o paciente. Nela, o médico conversa com o paciente, e tenta detectar indicadores e eventos que possam apontar para o diagnóstico correto.

Essa entrevista tem como objetivo detectar anomalias no processo de formação e expressão de idéias do paciente, como o esquecimento de detalhes importantes, ou a maneira de expressar seus sentimentos de maneira compreensível e razoável.

Atualmente esta análise é feita de forma completamente subjetiva pelo especialista, que através de sua entrevista(às vezes nem gravada) deve identificar, relatar e processar os indicadores achados num diagnóstico conclusivo, diagnóstico este que terá grande impacto na vida do paciente e de seus familiares. O problema aqui vem na variância de opiniões de médico para médico. Um médico pode ser parcial a um diagnóstico de demência para um paciente, enquanto outro a partir de uma nova consulta pode laudar um diagnóstico de epilepsia. Esta variação não decorre de uma má definição de características para as doenças, que são extremamente bem categorizadas em grupos de sintomas mutuamente exclusivos no Manual Técnico de Diagnóstico para Patologias Mentais(DSM-V), mas sim de uma dificuldade na interpretação de eventos decorrentes durante uma entrevista, e da variância em indicadores aparentes de entrevista para entrevista.

Outro fator que exacerba tal problema é a dificuldade de se auditar tal diagnóstico. Enquanto outras áreas da medicina tem seus diagnósticos constantemente auditados através de autópsias e revisões de laudos e sintomas clínicos, o mesmo processo se torna incrivelmente difícil quando há incompetência (propositalou não) de profissionais médicos. O registro de entrevistas, mesmo quando gravadas, é tida em segundo lugar à experiência do médico presente, devido a questões de fidelidade do conteúdo frequentes em um ambiente hospitalar, e matéria cefálica se deteriora excepcionalmente rápido após o óbito, o que torna pós-diagnóstico através de autópsia difícil e pouco conclusivo.

Dado estes fatores, se torna clara a necessidade de padrões de diagnóstico mais precisos e discretos, que possam facilitar tanto o diagnóstico precoce de doenças mentais, como a auditoria e a discussão de casos passados sob uma perspectiva mais estatística e clara tanto aos profissionais quanto aos pacientes e/ou seus familiares.

Para tal, defendemos a utilização de técnicas atuais de processamento de linguagem natural e de teoria dos grafos para que seja feita uma análise computacional de discursos de pacientes em busca dos mesmos indicadores definidos no DSM-V. Através de tais medidas, não só simplificamos o trabalho de médicos da área, como reduzimos consideravelmente a quantidade de incerteza envolvida em diagnósticos prematuros.

Este projeto é parte essencial de um projeto maior desenvolvido pela Liga Acadêmica de Neurociências com intenção de melhor correlacionar alterações na estrutura neuroanatômica do ser humano com alterações nos padrões normais de discurso. Através de tais correlações, esperamos poder gerar uma ferramenta de diagnóstico baseada apenas na coleta do discurso, o que irá simplificar em muito o diagnóstico de atipias mentais e poderá democratizar o acesso a tais diagnósticos.

A Seção 2 explicará como o projeto surgiu, e como foram coletados e processados os dados. Seção 7 mostrará um projeto feito com o mesmo objetivo que este, porém com uma abordagem diferente. Seções 4 a 6 explicarão as técnicas utilizadas no programa. A Seção 7 falará um pouco sobre a esquizofrenia, relatando seus sintomas e dificuldade de tratamento, e a Seção 8 relatará os resultados do nosso projeto, seguidos da conclusão e dos trabalhos futuros.

2. CONTEXTO

Esse trabalho surgiu no contexto do trabalho do aluno Gabriel Lins, um doutorando do Departamento de neurofisiologia do CCB da UFPE, que está sendo orientado pelo professor Marcelo Cairrão Araujo. Gabriel é o atual organizador da liga de neurociências, que se constitui num conjunto de palestras sobre neurociências para pessoas que estudam outros cursos na UFPE.

Essa liga, por criar e desenvolver idéias que se estendem para áreas diversas do conhecimento, é caracterizada por sua interdisciplinaridade.

Pelo fato do trabalho envolver um aluno de doutorado, a execução de sua idéia deve seguir alguns padrões impostos por um trabalho de doutorado, dentre eles, a idéia deve ser única e nenhum outro projeto publicado deve abordar a problemática de extrair métricas de pacientes da mesma maneira que este projeto aborda.

Importante ressaltar que o escopo do trabalho de doutorado, intitulado de "Análise diferencial de discurso em neuropatologias por modelagem computacional e diagnósticos por imagem" envolve muitos outros aspectos, como visualização de dados e análise de áudio, mas este trabalho irá abordar apenas um escopo limitado do trabalho de doutorado.

Esse projeto visa analisar traços de doenças mentais em pacientes, em tratamento ou não. A princípio, a própria liga de neurociências e afiliados pretendem utilizar a aplicação para avaliar sua eficácia com pacientes reais, e dependendo dos seus resultados, outros centros médicos podem chegar a utilizar essa aplicação.

2.1 BASE DE DADOS

Os dados dos pacientes com esquizofrenia foram concedidos pelo Ipub, e eles são o relato transcrito de entrevistas feitas com esse pacientes. Os pacientes, antes de serem entrevistados, passaram por um treinamento usando um programa que possui exercícios para melhorar o foco do paciente.

As perguntas usadas nas entrevistas estão no apêndice deste trabalho, e elas são divididas em 4 partes: motivação extrínseca, motivação intrínseca, desempenho e crença. A primeira parte da entrevista busca saber o que levou a pessoa a descobrir sobre o treinamento e que fator externo a motivou a fazer parte dele, como ajuda de custo ou recomendação de amigos. Já a parte de motivação intrínseca pergunta a respeito do que motivou a pessoa a continuar fazendo o treinamento e o que ela mais/menos gostou do programa.

A parte de desempenho fazia perguntas gerais a respeito do desempenho do paciente no programa, e a parte de crença pergunta sobre o quanto o paciente achava que esse programa o ajudou no tratamento da doença.

Um fator importante a ressaltar é que em cada entrevista havia uma observação a respeito do paciente, explicando o comportamento dele durante a entrevista e o período que ele terminou o treinamento, como no exemplo a seguir:

"[Participante], realizou treinamento no Ipub , é paciente do Hospital dia. Participante terminou o treinamento dia 16/10/2014 e foi entrevistado dia 28/11/2014. Participante demonstrou contentamento com o treinamento e feliz em poder falar sobre o seu feito. "

Algumas observações relataram que o paciente estava ansioso para o término da entrevista, ou que o paciente desvia do assunto e não consegue responder a pergunta apropriadamente. Essa última observação foi relativamente comum, pois já que a esquizofrenia dificulta o processamento de informações, é de se esperar que algumas pessoas continuem a apresentar esses sintomas mesmo depois do treinamento.

Essas entrevistas foram de fundamental importância para o desenvolvimento desse trabalho, embora o seu número ainda seja pequeno (foram 23 pacientes com esquizofrenia e 12 pacientes de controle)

2.2 PROCESSAMENTO DE DADOS

Com o acesso as dados, o programa executa duas tarefas principais: extrair métricas e treinar a IA. Para a primeira parte, o programa terá acesso a alguns componentes ajudar nessa extração, que são: Um corpus léxico adaptado para o português e o uso de bibliotecas como o nltk, numpy, spatial e networkx. O corpus léxico ajudará a analisar o sentimento de cada frase, a biblioteca networkx facilitará operações com grafos, o nltk ajudará a separar palavras e expressões de maneira mais precisa e as bibliotecas numpy e spatial ajudarão com cálculos envolvendo vetores de várias

dimensões. Com essas ferramentas, o programa consegue transformar um discurso em diversos grafos menores e calcular, entre outras métricas, a densidade e o diâmetro de cada um dos grafos, assim como, usando uma cópia inalterada do discurso, calcular a variação emocional de uma sentença a outra e obter com isso a variação emocional de cada parágrafo.

Em seguida, essas métricas são inseridas em um arquivo .txt no formato de arrays de floats. Cada array possui 16 números, sendo os 3 últimos para relatar o estado do paciente: se os 3 números forem, nessa ordem, 1,0,0; significa que o paciente é saudável, caso sejam 0,1,0; o paciente tem esquizofrenia. O terceiro número será útil quando o programa expandir e inserir uma terceira doença para análise, mas para fins do escopo desse programa, ele não tem utilidade.

Em seguida, uma segunda classe é iniciada e coleta esses dados para iniciar o uso da classificação de dados. Essa classe, assim que organizar os dados obtidos, vai separá-los em conjunto de treinamento e de teste, definindo a primeira das 3 últimas variáveis como alvo de predição, e usará 8 tipos de classificadores diferentes, que serão listados na Seção 4. Esses classificadores serão avaliados pela sua acurácia.

Esse resultado foi depois analisado e suas métricas (média e intervalo de confiança) foram apresentadas em um gráfico para posterior análise e conclusão dos resultados. Dessa maneira, foi possível analisar os discursos e retirar um significado relevante dos dados obtidos.

Um aspecto que destaca esse trabalho é a sua abordagem envolvendo assuntos tão distintos, o que oferece uma nova abordagem para o diagnóstico de doenças mentais e que pode ajudar bastante médicos envolvidos nessa área. Para exemplificar essa diferença, será analisado um exemplo de trabalho nessa área e explicadas as diferenças entre esse trabalho e este.

3. LITERATURA ATUAL

Um trabalho que tem aspectos semelhantes a esse, com o título de "Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls", feito por BERTOLA(2014), buscou analisar a fluência verbal entre pacientes saudáveis, pacientes com Alzheimer e pacientes com leve comprometimento cognitivo.

Para isso, foram incluídos cem idosos do Centro de Referência à Saúde do Idoso. Todos esses pacientes passaram por um teste clínico e neuropsicológico. O teste neuropsicológico consistiu em nove diferentes testes, incluindo o Mini Mental State Exam, um teste bastante utilizado e validado internacionalmente para esse uso. Para que o paciente fosse considerado com comprometimento cognitivo, ele deveria apresentar sinais desse comprometimento em pelo menos dois domínios cognitivos.

Em seguida, os pacientes foram divididos em quatro grupos, cada um com 25 membros: um grupo de pessoas saudáveis ou de controle, um de pessoas com problemas cognitivos em um domínio, um de pessoas com problemas cognitivos em dois domínios, e um de pessoas com Alzheimer.

Depois da divisão, esses pacientes fizeram um teste, que consistia em pedir para que os pacientes escrevessem a maior quantidade possível de nomes de animais, tentando evitar repetição. Em seguida, eles anotavam a quantidade de palavras corretas, de palavras erradas e de repetições. As respostas fornecidas pelos pacientes foram transformadas em grafos usando o programa *SpeechGraphs* (Mota et al., 2014).

Com o grafo obtido, eles calcularam a quantidade de palavras no grafo, além de outros 13 atributos: total de nós, total de arestas, maior componente fortemente conectado, maior componente fracamente conectado, arestas repetidas, arestas paralelas, ciclos de tamanho um, dois e três, grau médio, densidade, diâmetro, média dos caminhos mais curtos e coeficiente de clustering.

Como eles esperavam, o grupo de controle produziu mais nós, a rede produzida pelo grafo tinha um diâmetro maior e era menos densa, quando comparado com os outros grupos. Os grupos com problemas cognitivos obtiveram um desempenho intermediário entre o grupo de controle e o grupo com alzheimer em todas as medidas.

Quase todos os pacientes conseguiram evitar repetições nos nomes, porém 20% dos pacientes com Alzheimer repetiram alguma palavra com apenas duas palavras de distância (ex.: cachorro-gato-

cavalo-cachorro). Pacientes com Alzheimer tem problemas de processamento de memória desde os estágios iniciais, o que provavelmente explica a repetição de palavras em um intervalo tão curto.
Quanto ao nosso trabalho, existem algumas diferenças em comparação com este. A base de dados não foi produzida a partir de um teste de escrever nomes de animais, mas sim com entrevistas semi-estruturadas ou relatos de cuidadores de pacientes, tais cuidadores sendo saudáveis.

Também será analisado, além dos aspectos do grafo, a variação emocional dos discursos, que não seria possível caso o teste não permitisse uma expressão de sentimentos por parte do paciente. A variação emocional se mostrou como um forte diferenciador entre os tipos de pacientes, como será visto na Seção 8.

Além disso, a doença abordada foi diferente: enquanto o projeto de Bertola(2014) analisou pacientes com Alzheimer e com problemas cognitivos, enquanto este trabalho analisa especificamente pacientes com esquizofrenia, que possuem problemas mentais diferentes dos pacientes com Alzheimer, o que se reflete em tipos de discurso diferentes.

4. INTELIGÊNCIA ARTIFICIAL

Inteligência artificial é uma área do conhecimento que busca criar e desenvolver modelos computacionais que executem tarefas que envolvem tomadas de decisões complexas a fim de alcançar um objetivo, como vencer um jogo de xadrez ou reconhecer um objeto numa imagem.

Um software que use um modelo computacional dessa área obtém, por meio de treinamento usando uma base de dados, conhecimento a respeito das possíveis decisões a serem tomadas na resolução de um determinado problema, e usando esse modelo, esse software pode "aprender" a tomar as decisões certas.

Inteligência artificial possui muitas aplicações. Por exemplo, um site de filmes e séries pode utilizar uma rede neural para aprender os tipos de filmes que um determinado usuário prefere e recomendar novos filmes baseado na semelhança com os filmes anteriormente assistidos.

Um exemplo bastante famoso é o Deep Blue, um software criado pela IBM para jogar xadrez.

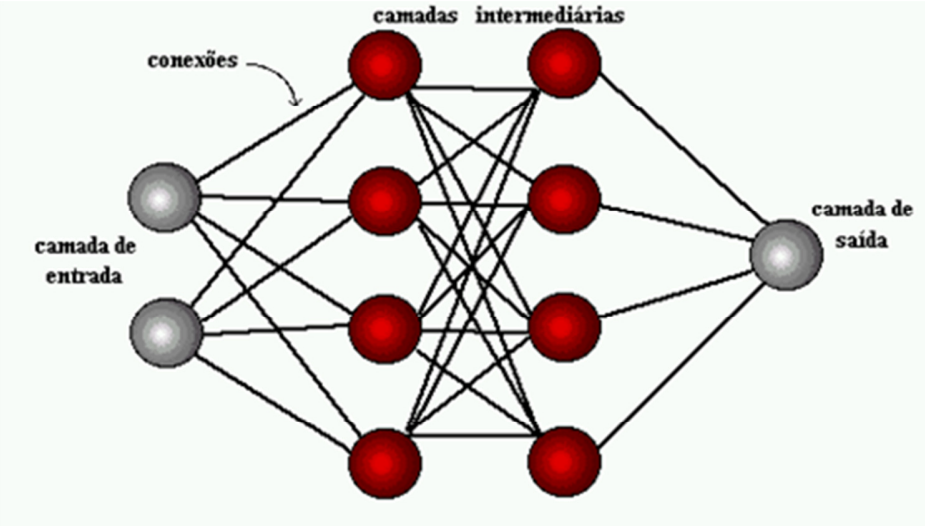
Segundo cad (2017), Deep Blue foi criado em 1985 por Feng-hsiung Hsu e em 1989 ele ganhou o nome atual. Ele derrotou o melhor jogador de xadrez do mundo, Kasparov, em maio de 1997, antes de ser aposentada pela IBM. Kasparov disse após a partida que "é verdade, me assusta que uma máquina jogue com um nível superior ao de todos os outros computadores e ao de quase todos os homens".

Existem diversos exemplos de técnicas para obter um bom classificador de dados. No momento, será explicado como funciona uma rede neural, um modelo computacional bastante conhecido.

A rede neural de um software como o Deep Blue tipicamente analisa todas as jogadas possíveis dentro de um escopo, feito por uma busca em profundidade limitada, e usando algoritmos de otimização de jogadas, como o min-max, ela calcula, considerando que o seu oponente também fará a melhor jogada possível, qual jogada mais a beneficia a longo prazo.

Uma rede neural tipicamente possui dois componentes: um conjunto de neurônios e um conjunto de ligações entre eles. As ligações entre neurônios, ou conexões, possuem pesos, e esse pesos indicam a relevância de um determinado fator na tomada de decisão daquele neurônio em particular.

Figura 1 - Organização em camadas



Fonte: DIN-UEM(2018)

São os pesos das conexões que irão moldar a rede neural a longo prazo. Cada neurônio da rede recebe informações vindas ou do ambiente externo (conjunto de dados) ou de outros neurônios, e ao dar pesos a essas informações, que podem ser positivos ou negativos, e tão grandes quanto o programa usado permitir, os neurônios posteriores receberão resultados moldados por esses pesos, resultados que se propagam até a camada de saída, que por fim exibirá o resultado final.

Esse resultado final será comparado com o resultado previsto para que os pesos sejam devidamente ajustados e medidas como precisão e acurácia sejam coletadas e analisadas.

Uma rede neural, para aprender um determinado conjunto de dados, ela tipicamente lida com 3 tipos de dados: dados de treinamento, de validação, e de teste.

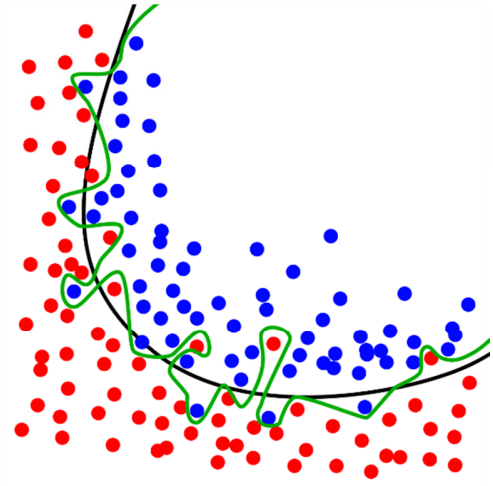
Nos dados de treinamento, a rede ajusta os pesos de cada conexão com o objetivo de aumentar sua taxa de acerto. Caso o resultado esteja errado, os pesos são ajustados.

Nos dados de validação e teste, não há ajuste de peso, mas há uma avaliação de desempenho. A rede neural compara seus resultados com os dos dados de validação geralmente a cada 5 iterações pelos dados de treinamento, embora esse valor possa ser alterado. Apenas quando se nota que a rede possa estar fazendo overfitting é que ela interage com os dados de teste, onde ela obtém o resultado final do seu desempenho.

Overfitting é quando uma rede neural tem sua função próxima demais de um determinado conjunto de dados. Uma analogia ao overfitting é o exemplo de um aluno que, ao invés de tentar aprender um determinado assunto, simplesmente decora todas as páginas do livro e sabe “de cor” o

que tem em cada página, mas não consegue explicar com suas próprias palavras o que ele entende de determinado assunto do livro.

Figura 2 - Overfitting



Fonte: en.wikipedia.org

Na figura acima, a linha em verde representa uma função de uma rede neural com overfitting, enquanto que a linha preta representa uma função de uma rede sem overfitting. Enquanto que a linha verde divide corretamente 100% dos dados e a linha preta comete alguns erros, caso um novo dado seja introduzido, há uma probabilidade maior de esse novo dado ser classificado corretamente pela linha preta do que pela linha verde, pelo fato de que a função representada pela linha preta está mais próxima da realidade dos dados do que a função da linha verde.

Uma rede neural possui pelo menos dois componentes: Uma camada de entrada e uma camada de saída, ambas compostas por neurônios. Os neurônios de qualquer camada recebem os dados iniciais, processam os dados, geram um peso baseado nos resultados e transmitem esse resultado à próxima camada. Se a camada de saída exibir um resultado diferente do previsto, a fórmula usada para processamento dos dados é alterada com base na taxa de aprendizado e na diferença entre o resultado previsto e o resultado obtido, usando um algoritmo chamado de Backpropagation.

Um grande problema ao se treinar redes com múltiplas camadas, segundo Deneck (2009), é que não se tem um resultado alvo para os neurônios da camada intermediária. Enquanto que a camada de saída pode comparar seus resultados com as saídas previstas, as camadas intermediárias não possuem tal parâmetro, porém elas também precisam ser ajustadas. Por esse motivo é que o

Backpropagation, que foi inventado nos anos 70 e usado com frequência em 1986, se tornou tão popular.

Backpropagation, que é uma abreviação de “backwards propagation of error”, é um algoritmo de aprendizagem supervisionada de redes neurais que usa gradiente descendente para ajustar os pesos dos neurônios de uma rede neural. (DENECK, 2009)

A inspiração para essa técnica surgiu do próprio cérebro, que busca naturalmente ajustar suas ações ao meio em que se encontra e com isso melhorar suas habilidades. O Backpropagation é aplicado a uma rede neural quando ela está em sua fase de treinamento.

Esse algoritmo necessita de uma função de erro para calcular o ajuste dos pesos dos neurônios em cada camada distinta. Essa é uma maneira de se ajustar corretamente os pesos dos neurônios da camada intermediária, pois uma vez que eles não produzem o resultado final mas apenas contribuem para ele, eles só podem receber o reajuste proporcionalmente a contribuição deles ao resultado final.

A parte regressiva do Backpropagation vem do fato de que a camada de saída é a primeira a receber o reajuste de pesos, enquanto que as outras camadas têm seus pesos ajustados por um fator reduzido da camada anteriormente ajustada pelo algoritmo.

Experts que examinaram redes de múltiplas camadas usando Backpropagation notaram que muitos nós aprenderam padrões similares aos que os próprios seres humanos detectaram durante a análise dos dados (E. DELISI, 2001). Por conta disso, um campo muito maior de áreas de aplicação de redes neurais se abriu, e a necessidade de experts para supervisionar o aprendizado em muitos casos deixou de existir.

Nesse sentido, uma rede neural tem vantagem sobre, por exemplo, uma árvore de decisão, já que uma árvore dessas precisa de um expert para definir as regras de decisão dela, enquanto que uma rede neural criará as próprias regras, que podem não ser compreendidas facilmente por um ser humano, mas serão altamente precisas na maioria dos casos.

Esse é o exemplo mais comum de inteligência artificial. Existem diversas outras maneiras de se classificar dados, e as maneiras que esse trabalho explorou serão detalhadas a seguir.

4.1 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL USADAS

Para esse trabalho, foram utilizados 8 classificadores diferentes. A razão pelo qual esse classificadores foram usados foi por suas diferentes abordagens ao problema em questão, além de

suas vantagens e desvantagens lidando com diferentes tipos de dados, o que pode revelar informações importantes a respeito da distribuição dos dados analisando apenas o desempenho dos diferentes classificadores. Além disso, usando uma regressão logística no classificador com melhor resultado, pode-se inferir a importância das medidas extraídas e reconhecer quais dados podem ser melhor aproveitados ou analisados.

Esses classificadores serão apresentados a seguir.

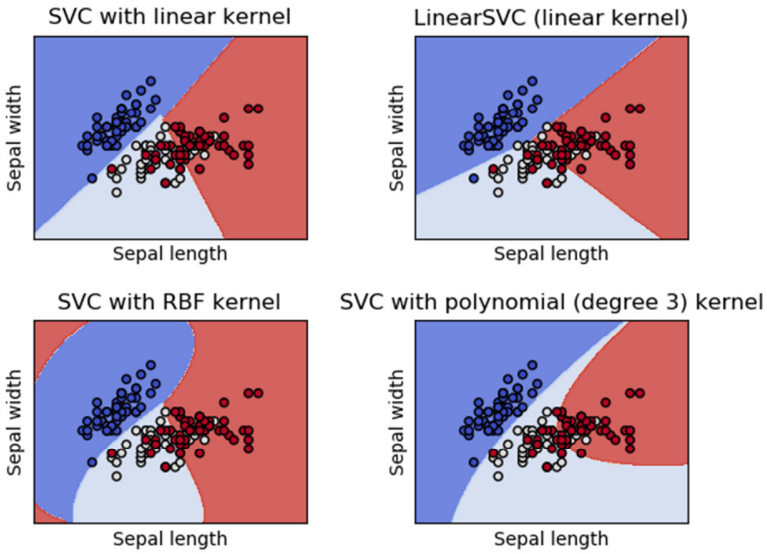
K Nearest neighbors: O método nearest neighbors pode ser feito de maneira supervisionada ou não-supervisionada. Ele memoriza todo o conjunto de treinamento e, quando é apresentado com um dado novo, ele analisa os K vizinhos mais próximos dele e o que a maioria decidir será o valor desse dado.

Nos dados dos discursos, os vizinhos são os vetores dos discursos que possuem a menor distância euclidiana em relação ao dado que quer ser analisado.

Essa técnica é boa em lidar com problemas multi-classe e é simples de implementar. Por outro lado, ele precisa de uma função de calcular distâncias entre dados que seja relevante e também ignora toda a base de dados para decidir o valor de um dado novo baseado apenas em um subconjunto de dados próximos a ele.

Support vector classifier(SVC): Esse classificador utiliza uma série de métodos de aprendizagem de máquina para classificar dados. Qual(is) elemento(s) da série será(ão) usado(s) depende dos parâmetros aplicados. Na figura abaixo, temos 4 exemplos de como o SVC pode separar o mesmo conjunto de dados quando usado com parâmetros diferentes.

Figura 3 - Representação de modelos SVC



Fonte:scikit-learn.org

Os dois primeiros exemplos usam de classificação linear, porém as bibliotecas usadas são diferentes(libsvm e liblinear). O terceiro exemplo usa kernel RBF, que usa de classificação radial para separar os conjuntos de dados. O último usa classificação polinomial, e com isso ele consegue produzir uma curva descrita por uma função quadrática para separar os conjuntos.

SVCs são eficientes em conjuntos de muitas dimensões e, por usarem vetores de suporte para previsões, também são eficientes em custo de memória. Por outro lado, podem facilmente cometer overfitting quando usam dimensões demais para classificar.

Gaussian: Esse classificador utiliza da distribuição gaussiana para agrupar os dados do conjunto de treinamento. Com esse agrupamento, os dados são separados em clusters, que é um elemento que possui um centróide e uma forma, que pode ser elíptica ou redonda. Quando esse classificador recebe um novo dado e deve classificá-lo, ele analisa que cluster possui sua borda mais próxima do elemento em questão e então classifica esse elemento como parte desse cluster. O cluster possui uma identificação de antemão, como “cluster de controle” ou “cluster de esquizofrenia”, para poder classificar os dados novos.

Árvore de decisão: Uma árvore de decisão utiliza as ramificações de sua estrutura de árvore para criar um fluxo de análises e tomar decisões a partir dela. A fase de treinamento consiste em criar

os ramos e aprimorar os critérios de cada ramo de maneira a otimizar a divisão entre classes do conjunto. No fim, as folhas representam a classe em que o dado será inserido.

Uma árvore de decisão tem uma boa performance quando os resultados são discretos e quando o conjunto tem um desequilíbrio na proporção de dados, o que torna árvores de decisão um bom classificador para detectar fraudes de cartão de crédito, por exemplo. Ela também tem uma boa maneira de expressar sua classificação, uma vez que precisa apenas exibir os critérios usados em cada ramo para tomar decisões.

É importante evitar o overfitting em uma árvore de decisão, e para isso, as técnicas de podar ramos da árvore devem ser bem aplicadas. Além disso, uma árvore ótima para o problema teria uma complexidade NP-completa, o que pode ser um obstáculo em termos de desempenho.

Random Forest: O método random forest consiste em criar diversas árvores de decisão usando o método de bagging e combinar suas predições para formar um único classificador. O método de bagging é um método estatístico que busca melhorar a estimativa dos seus elementos combinando a estimativa de vários outros elementos. Dessa maneira, cada nova árvore de decisão será melhor em classificar os elementos que as anteriores não conseguiram classificar bem. Essa abordagem evita o overfitting tão comum em árvores de decisão, e reduz a complexidade de cada árvore do classificador.

Em geral, random forest é simples e rápido de usar, e pode ser usado tanto para regressão quanto para classificação de dados, além de ser difícil de ocorrer overfitting nesse algoritmo. Por outro lado, o random forest pode se tornar muito lento se a quantidade de árvores se tornar muito alta, o que depende também da capacidade de processamento do computador.

MLP: O tão conhecido multi-layer perceptron. A maior parte do que foi explicado neste capítulo sobre inteligência artificial se aplica ao MLP, uma vez que o próprio Deep blue é um exemplo de MLP. Ele utiliza os conceitos de backpropagation e feedforward em uma rede neural de múltiplas camadas para aprimorar os pesos das conexões presentes com o objetivo de melhorar a classificação do conjunto de dados de validação e teste.

GaussianNB: Esse classificador usa o método naive bayes para classificação. Ele usa métodos estatísticos simples e converte os dados em uma tabela de frequência para calcular as probabilidades de cada evento influenciar a resposta ou não. Ele também usa as fórmulas de uma distribuição gaussiana para obter a função de densidade probabilística e estimar como os dados se distribuem no espaço vetorial.

Figura 4 - Exemplo de tabelas usadas pelo modelo GaussianNB

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Fonte: analyticsvidhya.com

No exemplo acima, O conjunto de treinamento cita o tempo em determinado dia e se uma determinada pessoa jogou tênis ou não. Note que a tabela a seguir mostra as frequências relacionadas com os seus resultados. Quando o tempo era “overcast”, por exemplo, em 100% das vezes essa pessoa jogou tênis, o que só aconteceu em 60% dos casos em que o tempo era “sunny”. Em seguida, de maneira intuitiva, calculam-se as probabilidades de acontecer cada evento separadamente, como a chance de ter um dia “rainy” ou a chance dessa pessoa jogar tênis em um dia aleatório. Por fim, usando o teorema de Bayes, ele calcula qual a chance dessa pessoa jogar tênis dado que está “sunny”, por exemplo, e caso a probabilidade esteja acima de 50% ele classifica o dia como um “dia de tênis”. Esse classificador tem uma boa performance em casos em que as variáveis têm pouca correlação e o GaussianNB é rápido e simples de implementar. Por outro lado, caso as variáveis sejam relacionadas, ele terá uma baixa performance.

QDA: Esse é o classificador quadrático de bayes. Enquanto que outros classificadores, como LDA, usam uma função linear para classificar os dados, o QDA usa uma função quadrática, por conta da função de probabilidade utilizada. Diferentemente do GaussianNB, esse classificador assume que os dados estão presentes em uma distribuição normal e portanto, usa fórmulas dessa distribuição para estimar a posição de dados novos e assim classificá-los.

Apesar de ser relativamente simples de implementar, o QDA usa sempre uma função quadrática para separar os dados, o que se torna o seu defeito mais notável, além do overfitting que pode ocorrer como consequência de tentar classificar dados distribuídos de maneira linear em uma função quadrática.

5. ANÁLISE DE SENTIMENTOS

Informação textual pode ser categorizada em fatos e opiniões. Fatos são expressões objetivas a respeito de entidades ou eventos. Opiniões são expressões subjetivas descrevendo atitudes ou sentimentos a respeito de entidades. De certo modo, o conceito de sentimento ou opinião é relativamente vasto, pois inclui subjetividade, polaridade, emoções ou até comparações (DENECKE; DENG). Enquanto que, por exemplo, dizer se gostou ou não de um filme, show ou qualquer outro evento é algo simples e direto, capturar o sentimento principal em um comentário é um desafio mesmo nos dias atuais.

Pesquisas em análise de sentimentos começaram em 2004 (DENECKE; DENG) , e desde então vem sendo usada em diversas áreas do conhecimento, desde o cinema até a medicina. Uma de suas aplicações é diferenciar um comentário positivo de um negativo, como por exemplo, saber se o quadro clínico de um enfermo melhorou ou não, analisando apenas os comentários sobre seu estado de saúde desde que foi atendido pela primeira vez.

O uso das técnicas relacionadas a essa área permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações, regras e realizar análises qualitativas e quantitativas em documentos de texto (NUNES, 2016). Tamanha utilidade permitiu que ferramentas mais rebuscadas fossem introduzidas nessa área. Se nos seus primórdios, essa análise apenas classificava um comentário como positivo ou negativo, com o tempo, tornou-se possível categorizar o sentimento específico presente no comentário e a sua intensidade, que, sob uma análise mais objetiva, percebeu-se que pode ajudar a entender a personalidade do locutor.

A análise de sentimentos, além de ser usadas na otimização de diversas técnicas computacionais, também contribui com o avanço da medicina, porém, existem alguns problemas que quem quer que se interesse nessa área deve levar em consideração antes de produzir um categorizador de textos.

Um dos desafios nessa área envolve a interdisciplinaridade. Existem poucos pesquisadores que possuem conhecimentos em ambas as áreas de computação e biologia, e muitas aplicações na área biológica não são stand-alone: pelo contrário, são aplicações integradas e que dependem de uma coordenação eficaz para trazerem bons resultados.

Por conta disso, nota-se uma certa fraqueza dos pesquisadores em defenderem seus produtos, algo que não deveria acontecer, considerando a quantidade de vidas que foram salvas graças a essas tecnologias.

A granularidade da análise pode gerar resultados bem distintos e pode não ter a utilidade desejada pelo usuário. Um classificador que categorize um parágrafo inteiro, por exemplo, pode não exibir a diferença de sentimentos entre duas frases com sentimentos (ou polaridades) distintas, como "O filme foi muito bom, pena que o ingresso foi caro.". Mas um classificador que analise o sentimento de cada palavra em particular pode perder o senso de contexto e trazer dificuldades a um usuário que esteja apenas interessado no sentimento geral presente no texto todo, ao invés de estudar cada detalhe dele.

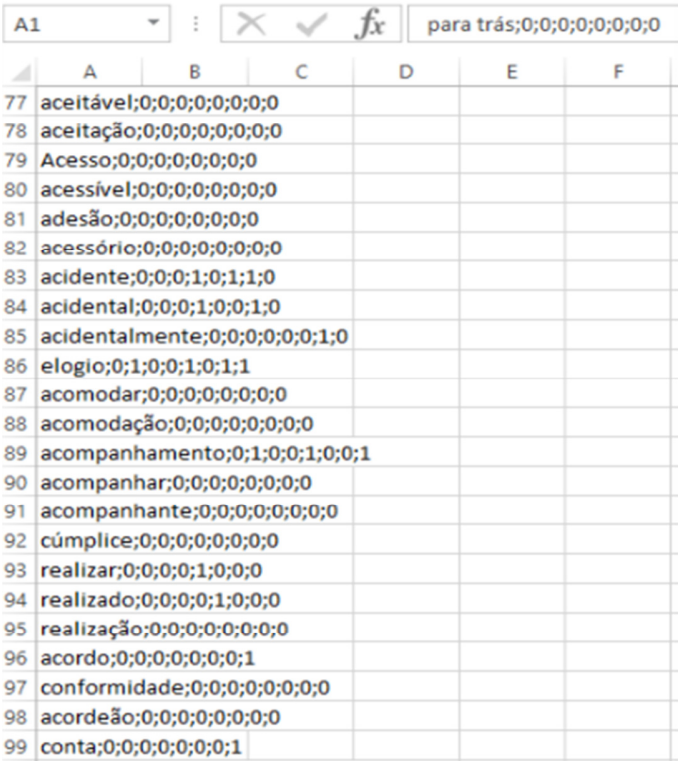
Além disso, alguns estudos perceberam que a análise de sentimentos, também conhecida como mineração de opiniões, costuma ser dependente de contexto (DENECKE,2015), uma vez que as palavras podem expressar um sentimento diferente dependendo do domínio em que a sentença, texto ou frase esteja inserida. Por conta disso, um classificador que usa análise de sentimentos deve ser treinado com uma base de dados de textos apenas daquele domínio específico, e seu vocabulário deve ser adaptado para a interpretação do contexto considerado.

A maneira mais direta de se aumentar a representatividade de um lexicon é estendendo-o a um domínio específico. Um lexicon é a representação do vocabulário ou área do conhecimento. No contexto de análise de sentimentos, um lexicon é o mapeamento de uma palavra em um vetor de sentimentos. Esses lexicons possuem um contexto em que eles podem ser aplicados corretamente, pois enquanto que um lexicon que represente o uso de uma palavra em um contexto geral não será de fato mais que isso, um lexicon de domínio específico pode ignorar o uso mais geral da palavra e cometer uma espécie de over-fitting. Nesse caso, a solução mais utilizada é fazer um merge do lexicon mais geral com o de domínio específico, evitando que quaisquer das partes do conhecimento da palavra seja ignorado ou desconsiderado.

De fato, muitas abordagens a análise de sentimentos são baseadas nos seus lexicons, já que eles são a unidade básica de sentimento nesse contexto. Existem diversas bases de dados de lexicons, muitas delas em inglês, como SentiWordNet e WordNetAffect (NUNES, 2016). Essas bases de dados armazenam os mapeamentos dos lexicons para que, usando algoritmos de predição de sentimento, possa-se intuir o sentimento predominante em determinada parte do texto.

Algumas dessas bases de dados são mais gerais, enquanto que outras são de um domínio mais específico (como medicina, ou política). Para o nosso projeto, foi usada a base de dados NRCEmotion, disponibilizada pela PUC de São Paulo. Essa base de dados consiste em um arquivo .csv com cada linha representando uma palavra e o valor dos sentimentos associados a esta palavra. A figura abaixo mostra uma parte desse arquivo:

Figura 5 - Trecho de corpus léxico da USP



Fonte: O autor(2018)

A aparente enorme quantidade de zeros nos valores se deve ao fato de que diversas palavras possuem um cunho emocional pequeno ou nulo e portanto não denotam claramente nenhum sentimento em específico. Os 8 valores que seguem cada palavra ou expressão representam cada um desses sentimento, em ordem: Raiva, antecipação, desgosto, medo, alegria, tristeza, surpresa, confiança.

Outro desafio nessa área envolve a criação e interpretação de diagnósticos. Há uma transição do tradicional método médico-paciente de diagnosticar para o modelo baseado em guidelines. Essa transição ocorre porque, com o advento da inteligência artificial, o médico deve não só interpretar os sinais e sintomas do paciente, como também deve entender o resultado que a sua aplicação oferece, aplicação esta que pode usar redes neurais para chegar a uma conclusão, que é um método eficiente, mas o médico terá muitas dificuldades para entender quais pesos na rede neural foram responsáveis pelo resultado e que outros resultados tinham uma chance razoável de serem escolhidos, isso se o médico entender a rede. Desse modo, os pesquisadores tiveram um outro desafio pela frente, que é o

de usar a ontologia na representação do conhecimento de maneira que seus resultados, assim como seus processos, sejam compreensíveis ao médico que utilizará o software.

Uma maneira de se contornar essa situação é fazendo uma análise de dados de forma inteligente. A inteligência artificial pode oferecer um diagnóstico preciso, mas um estudo profundo da base de dados pode oferecer conhecimento tácito para se permitir uma apresentação explícita do conhecimento utilizado para seus diagnósticos. Esse conhecimento pode, por sua vez, auxiliar na gestão do conhecimento e permitir a produção de aplicações mais robustas e com uma taxa de erros menor.

5.1 METODOLOGIA EM ANÁLISE DE SENTIMENTOS

Para realizar o cálculo de variação emocional, não foi feita nenhuma remoção de stopwords e stemming no texto, uma vez que isso prejudica o desempenho da análise. Além disso, foi utilizado o corpus léxico mostrado anteriormente e a biblioteca Spatial, que lida com algumas operações vetoriais necessárias para esse trabalho, como calcular a distância euclidiana entre dois vetores.

Primeiramente, a entrada para esse método é um parágrafo inteiro, mas o resultado do cálculo de variação emocional é aplicado em cada frase do parágrafo. Isso acontece porque a granularidade dos outros métodos e técnicas é maior e lida com cada frase separadamente, porém se o mesmo fosse feito com análise de sentimentos, muito do contexto seria perdido. Para conciliar essa diferença, o cálculo dessa variação se mantém sobre o parágrafo inteiro, mas o resultado é distribuído entre suas frases.

Usando o parágrafo como entrada, o programa tokeniza o discurso do paciente em frases, utilizando o tokenizador padrão do NLTK, e então tokeniza cada uma das frases individualmente, gerando assim uma matriz de palavras. Após feito isso, para cada frase, o programa busca em todo o léxico por palavras ou expressões que estejam contidas nessa frase e adiciona o seu valor emocional à frase, representada por um vetor de 8 dimensões, um para cada emoção: Raiva, antecipação, desgosto, medo, alegria, tristeza, surpresa, confiança. Ao final desse processo, cada frase terá um vetor de emoções associado. Uma vez feito isso, a última parte do programa calcula a distância euclidiana do vetor emocional de uma frase para o vetor da frase seguinte usando a biblioteca Spatial, gerando uma lista com todas essas distâncias. Por fim, calcula-se a média entre essas distâncias, e essa média será o valor aplicado a cada frase.

O motivo pelo qual a variação emocional foi escolhida deve-se a uma análise detalhada dos sintomas de esquizofrenia. Um paciente esquizofrênico possui uma certa instabilidade emocional em seu discurso devido às suas constantes fugas de tema (também chamadas de para-respostas), o que torna difícil diagnosticar um paciente baseado apenas no sentimento demonstrado em um parágrafo ou mesmo a intensidade desse sentimento, porém torna relativamente simples fazer esse diagnóstico observando a variação de sentimentos entre frases.

6. TEORIA DOS GRAFOS

Grafos estão presentes em diversos lugares, e sua representação gráfica simples de interpretar torna seu uso muito comum na sociedade. Por exemplo, o famoso metrô de londres possui um grafo representando o conjunto de paradas pelo qual ele passa. Muitos londrinos usam esse metrô frequentemente e não têm dificuldade alguma em interpretar aquele conjunto de pontos ligados por arestas, com cada ponto tendo o nome da estação correspondente acima.

No caso anterior, uma aresta entre dois pontos representa a relação “Depois da estação X, o metrô segue para a estação Y”. Em um grafo que represente uma árvore de decisão, a relação representada é “Depois de tomar essa decisão, siga para essa pergunta”, e num discurso de um paciente, uma aresta representa “Depois da palavra X, o paciente falou a palavra Y”, e cada um desses grafos, mesmo com uma representação gráfica semelhante, possui usos diferentes e características relevantes diferentes.

Grafos, de uma maneira geral, simplificam relações complexas com ícones de fácil compreensão. Um grafo é um conjunto de objetos chamados de vértices, ligados por arestas. Esses vértices e arestas podem ter diversas propriedades, além do próprio grafo possuir determinadas métricas.

O uso de grafos normalmente simplifica a visão do cenário em questão, mas também atribui propriedades relevantes a ele, e isso é estudado pela teoria dos grafos.

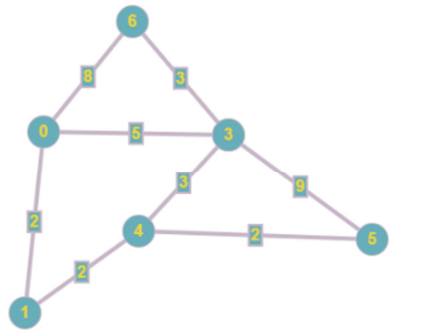
Em ciências da computação, a teoria dos grafos é o estudo dos grafos, uma estrutura matemática que relaciona objetos em pares, representados por nós e vértices.

Um grafo direcionado possui vértices, arestas e um mapa s, t : E → V, onde s(e) é a fonte e t(e) é o alvo da aresta direcionada “e”. Dessa maneira, sua representação é ligeiramente diferente da

de um grafo não-direcionado, pois as arestas de um grafo direcionado possuem setas indicando a direção que elas seguem.

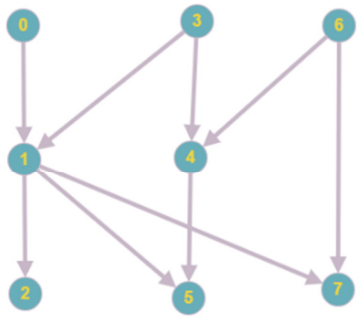
Quando um grafo possui arestas com peso, a sua representação também sofre mudanças adequadas para representar esse novo dado.

Figura 6 - Grafo com pesos



Fonte: O autor(2018)

Figura 7 - Grafo direcionado



Fonte: O autor(2018)

Na figura 6, suas arestas têm peso, mas permanecem não-direcionadas. Já na figura 7, suas arestas não têm peso, mas são direcionadas. Um grafo simples não possui pesos ou direções nas arestas.

Teoria dos grafos pode ser usada em mineração de dados, segmentação de imagens, agrupamento, captura de imagens, análise de redes, etc. (RIAZ; M. ALI, 2011) . Uma questão tipicamente abordada pela perspectiva de grafos é o problema do caixeiro-viajante. A representação desse problema é feita com um grafo não-direcionado e com pesos nas arestas, onde os nós são as cidades e os pesos dos vértices são as distâncias entre as duas cidades ligadas por ele.

Os conceitos a respeito de grafos que serão utilizados durante este trabalho estão representados a seguir.

Grafo simples: É um grafo onde suas arestas não possuem direção, sem laços e sem arestas paralelas.

Subgrafo: O subgrafo de um grafo qualquer é um grafo cujo conjunto de vértices e de arestas é um subconjunto do grafo original.

Grafo direcionado: É um grafo onde suas arestas possuem direção.

Densidade: É a proporção de arestas por vértice em um grafo.

Ciclo: É um subgrafo em que cada par de nós está ligado através de um conjunto de arestas. O grau de um ciclo refere-se a quantidade de nós presentes no subgrafo, e um ciclo de grau 1 é um subgrafo com 1 nó que possui uma aresta ligando o nó a ele mesmo.

Componente conectado: Em um grafo direcionado, é um subgrafo em que cada par de nós está ligado através de um caminho de arestas. Esse componente é uma versão mais restrita de um ciclo.

Componente fortemente conectado: Em um grafo direcionado, é um subgrafo no que cada par de nós é ligado mutuamente, havendo assim um caminho que ligue quaisquer pares de nós do grafo seguindo qualquer direção.

Aresta repetida: Caso duas arestas liguem o mesmo par de nós, essas arestas são consideradas repetidas.

Diâmetro: É a maior distância possível entre dois vértices quaisquer de um grafo.

Média do caminho mais curto: É o valor médio da distância entre dois pontos diferentes quaisquer do grafo.

Um problema que foi abordado pelo projeto foi a detecção de ciclos em um grafo. No nosso caso, cada nó do grafo representa uma palavra, e um vértice entre dois nós significa que uma palavra foi dita logo após a outra. Note que nesse caso, as arestas não possuem peso e são direcionadas.

Esses ciclos no grafo representam repetição de frases ou palavras num discurso de uma pessoa. Apesar de sermos incentivados a ter um vocabulário cada vez mais rico, e assim evitar a repetição de palavras, muitas vezes não temos criatividade o suficiente, ou queremos evitar o preciosismo, ou queremos enfatizar determinado aspecto de nossa fala, o que nos leva a gerar ciclos no grafo representativo. Busca-se obter um conjunto de métricas neste trabalho, entre elas ver com que frequência pessoas com determinados problemas mentais geram esses ciclos comparados com pessoas que não possuam esses problemas, qual os tamanhos desses ciclos, se esses ciclos possuem componentes completamente conectados entre si, e qual a densidade do grafo.

Um paciente com problemas mentais pode, por exemplo, não lembrar corretamente os detalhes importantes de alguma história que queira contar. Se for pedido a esse paciente que conte uma história envolvendo 3 ratos, ele pode se esquecer desse detalhe e incluir apenas 2 ratos na sua narrativa, ou pode citar os 3 ratos brevemente no começo da história, mas depois focar a história apenas em um dos ratos e ignorar completamente a existência dos outros dois. Imprevisíveis como são, doenças mentais podem produzir diversos efeitos na hora de contar uma narrativa simples, como descrever um sonho, ou contar algo que lhe aconteceu recentemente.

6.1 ALGORITMOS EM TEORIA DOS GRAFOS

Pelo fato de teoria dos grafos lidar com muitos problemas considerados como NP-completos, vários algoritmos podem ser usados para encontrar respostas próximas da exatidão em tempo hábil. Os problemas que esse trabalho busca solucionar com algoritmos em grafos são:

1. Encontrar ciclos de tamanho específico em um grafo.
2. Encontrar o maior ciclo presente em um grafo.
3. Encontrar componentes fracamente conectados e fortemente conectados em um grafo.
4. Encontrar a densidade de um grafo.
5. Encontrar arestas paralelas e laços em um grafo.
6. Encontrar a média do caminho mais curto em um grafo.
7. Encontrar o diâmetro de um grafo.

6.2 TÉCNICAS DE TEORIA DOS GRAFOS USADAS

Para esse trabalho, foi utilizada a biblioteca networkx de python, que é projetada para criar grafos e extrair propriedades deles. Na função responsável por extrair essas características, pode-se ver o seguinte código:

```
def searchCycles(dg):
        cycles = list(nx.simple_cycles(dg))
        …
```

A função nx.simple_cycles() retorna uma lista com todos os ciclos presentes no grafo passado como parâmetro, usando uma versão iterativa do algoritmo de johnson. Nessa mesma função, também há o seguinte código presente:

```
...
try:
        sCycles = list(nx.find_cycle(dG, orientation='original'))
        strongComp =max([len(s) for s in sCycles])
...
```

Nesse caso, a função nx.find_cycle() procura por ciclos mas leva em consideração que o grafo é direcionado, e com isso a ordem dos nós em cada aresta é considerada o fator principal para a direção destas. Quando o parâmetro orientation é ‘original’, a direção da aresta vai do nó da esquerda ao nó da direita, quando é ‘reversed’; da direita para a esquerda. Esse dado foi usado para se obter o maior componente fortemente conectado do grafo, como foi feito na linha seguinte do código. Essas duas funções são exemplos de como foram extraídas as métricas necessárias dos grafos para o trabalho.

7. DOENÇAS MENTAIS ABORDADAS

A princípio, planejava-se abordar pelo menos 2 doenças a fim de não apenas poder comparar esses pacientes com pessoas saudáveis, mas também comparar esses dois grupos de pacientes entre si. Por falta de dados decidiu-se analisar apenas a doença de esquizofrenia.

7.1 ESQUIZOFRENIA

A esquizofrenia é uma doença crônica que geralmente atinge pessoas entre 16 e 30 anos. A esquizofrenia afeta como a pessoa pensa e age: uma pessoa com esquizofrenia aparenta ter perdido o contato com a realidade (NATIONAL INSTITUTE OF HEALTH, 1997).

Essa doença, além de possuir um fator genético associado, pode ser causada por uma anomalia na estrutura cerebral. Acredita-se que alguns hormônios, como dopamina e glutamato, também possuem um papel importante na formação, ou prevenção, da doença.

Os sintomas dessa doença costumam envolver o processamento e execução de informações, além de distúrbios no pensamento em geral. Esses sintomas incluem alucinações, dificuldade em lembrar informações, expressão emocional reduzida, problemas de atenção e dificuldades manter o foco em um assunto. De fato, durante as entrevistas realizadas com os pacientes, a maioria deles não conseguia manter o foco na pergunta e mudava de assunto completamente no meio da resposta.

Segundo National Institute of Health (1997), o distúrbio de linguagem é uma característica central da esquizofrenia e um dos principais comportamentos pelos quais ela é diagnosticada. A gramática fica razoavelmente intacta, mas o conteúdo pode se perder ou ser incoerente, um sintoma referido comumente como “afrouxamento de associações”. Padrões de discurso mais bizarros, porém menos comuns, incluem neologismos (palavras criadas de um modo peculiar), bloqueio (interrupções espontâneas repentinas) ou associação ressonante (associações com base nos sons e não nos sentidos das palavras).’

Além dos sintomas mentais, há também uma deficiência na coordenação motora, que surge com as falhas na comunicação cérebro-corpo. Essa deficiência é refletida com tremores nos membros superiores e inferiores, assim como uma certa dificuldade em articular a fala.

Por conta desses tremores, uma pessoa com essa enfermidade costuma ter muita dificuldade em realizar tarefas diárias simples, como se alimentar e se deslocar pela casa. Pessoas assim precisam de atenção constante, e geralmente são seus parentes (esposa, mãe, filhos) que resolvem passar mais tempo em casa para garantir que essa pessoa possa receber o tratamento adequado fora dos hospitais.

A esquizofrenia também se reflete na maneira de falar de um paciente. Os sintomas a seguir foram percebidos durante a análise dos discurso de pacientes com esquizofrenia:

Logorréia: A logorréia (ou verborreia, ou verborragia) refere-se à uma expressão verbal aumentada. O paciente fala o tempo todo e é difícil interrompê-lo. Ela é observada tipicamente nos estados maníacos.

Oligolalia: A oligolalia (laconismo) refere-se a uma expressão verbal diminuída, mas não abolida. Consiste no oposto da logorréia. É observada nas mesmas situações em que o mutismo pode ocorrer.

Para-respostas: são respostas disparatadas em relação às perguntas . Por exemplo: "Qual é o seu nome?" – resposta: "acho que vai chover hoje!". Este distúrbio ocorre na esquizofrenia e na demência.

Neologismos: consistem em palavras novas, criadas pelos pacientes, ou palavras já existentes às quais é atribuído um novo significado. São encontrados principalmente na esquizofrenia.

Esses sintomas possuem suas ressalvas quanto a precisão de detecção de uma doença. A logorréia, por exemplo, apesar de ocorrer tipicamente em pacientes com esquizofrenia, pode ser presente como sintoma de outras doenças e, portanto, pode não ser definitivo em um diagnóstico.

Todos esse sintomas podem acarretar em um impacto psicológico no paciente, uma vez que ele pode se sentir excluído da sociedade. Embora seja o papel do médico cuidar desses sintomas, o papel desse projeto é identificar os sintomas e para isso, ele deve levar em conta todas essas mazelas decorrentes da doença.

O tratamento da esquizofrenia envolve controlar os sintomas dela, uma vez que não se sabe ao certo a raiz do problema. Por isso, tratar a esquizofrenia normalmente envolve o uso de antipsicóticos e de tratamento psicossocial, que envolve a ajuda de um médico para auxiliar o paciente a ter uma vida normal e superar os obstáculos que surgem com o desenvolvimento da doença.

Espera-se que o discurso de uma pessoa com esquizofrenia normalmente reflita o desvio de assunto e a repetição de informações, proveniente da dificuldade de articulação da fala, associada com a falta de foco do paciente. Além disso, pelo fato de que um paciente com esquizofrenia normalmente se isola socialmente devido às dificuldades de locomoção e, estatisticamente

desenvolver alguma forma de depressão, espera-se notar uma variação emocional menor nesses discursos. Fazendo uma análise detalhada das ocorrências de ciclos nos grafos gerados, será possível detectar a repetição de informações, assim como uma falta de conexão na resposta.

8. RELATÓRIO

Foi elaborado um extrator de características do texto, que poderá ser utilizado nos discursos transcritos de pacientes atípicos e entregará características relevantes destes aos médicos, a fim de facilitar um diagnóstico conclusivo com base em indicadores de discurso como previsto pelo DSM-V.

Neste trabalho, foram extraídas 8 características fundamentais do texto:

• Grau médio total(ADT), maior componente conectado(LLC), maior componente fortemente conectado(LSC), arestas paralelas(PE), ciclos de um nó(C1), ciclos de dois nós(C2), ciclos de três nós(C3), diâmetro(DM) e média dos caminhos mais curtos (SP) são extraídos através da criação e medida de uma representação de grafo para o discurso, e como discutido em trabalhos anteriores é essencial para o diagnóstico de psicoses leves e agudas.

• Variância emocional é uma métrica de curva que é tido como o quanto o cunho emocional de frases varia ao decorrer do discurso. Para extraí-la, usamos um léxico emocional para obter a soma do cunho emocional de todas as palavras de uma frase, e a distância destas somas frase a frase se torna nossa medida de variância no tempo.

O sistema será desenvolvido usando uma combinação de programas em java e em python, fazendo uso da biblioteca de processamento de linguagem natural NLTK.

Como entrada, o programa receberá discursos de pacientes propriamente formatados como definido na Seção Entrada/Saída, e como saída o sistema deve entregar as medidas obtidas através da análise desses discursos.

Finalmente, os testes deste sistema serão dados a partir de três objetivos necessários para o funcionamento de tal sistema.

• O sistema deve fornecer métricas determinísticas e replicáveis dado um discurso de paciente e um mesmo corpus de base.

• O sistema deve fornecer valor aos médicos da área, o suficiente para justificar seu uso em atividades diárias de diagnóstico.

• O sistema deve entregar resultados coerentes com os laudos de pacientes de teste.

O detalhamento de tais resultados é discutida com mais detalhes na Seção 8.3, Testes.

8.1 ENTRADA/SAÍDA

Como entrada, o programa recebeu 34 discursos transcritos em português, sem erros gramaticais ou semânticos, divididos em 23 textos de pacientes esqizofrênicos e 12 textos de pacientes de controle.

Como saída, o programa retorna um arquivo patient_data.txt, contendo as medidas de média e desvio padrão, nessa ordem:

1. Arestas Repetidas
2. ciclos de grau 1
3. ciclos de grau 2
4. ciclos de grau 3
5. Máximo componente conectado
6. Máximo componente fortemente conectado
7. Grau médio
8. Diâmetro
9. Média dos caminhos mais curtos
10. Tamanho da frase
11. Número de nós
12. Número de arestas
13. Variância emocional

Descrições dessas medidas estão presentes nas Seções 5 e 6.

8.2 FUNCIONAMENTO DO PROGRAMA

Ao receber o discurso do paciente em formato .txt o programa faz um pré-processamento para remoção de stopwords, e depois realiza stemming. Como queremos gerar um grafo direcionado de palavra a palavra carregada de sentido do texto, palavras sem relevância textual, como “o”, “a”, “como” e “não” levariam a uma super-avaliação destas palavras quando fossemos extrair as características do grafo, o que geraria ruído nos resultados. Realiza-se a filtragem de stemming pela

mesma razão, visando reduzir a inclusão demasiada de palavras devido às muitas conjugações pessoais e temporais da linguagem portuguesa.

Recursos adicionais: Para este processamento, foi utilizado o pacote de Python Networkx versão 2.1 (HAGBERG, 2018), que auxilia na manipulação de grafos. Além disso, para a análise de sentimentos, este programa utiliza um léxico emocional para extrair o conteúdo emocional de cada palavra do texto. Agradeço a instituição PUC de São Paulo, que forneceu o léxico NCR Emotion como léxico emocional, o que tornou possível a realização desse trabalho. Esse léxico consiste de um conjunto de cerca de 14 mil palavras e expressões seguidas de um vetor de 8 números, representando 8 sentimentos, em ordem: raiva, antecipação, desgosto, medo, alegria, tristeza, surpresa, confiança.

Método: O programa separa o texto em frases e calcula as medidas referentes ao grafo representativo de cada frase e a variação emocional do parágrafo que esta frase pertence, formando uma lista de dados. Em seguida, essas listas de dados são usadas por uma classe que utiliza de 8 classificadores usando inteligência artificial para treiná-los na identificação do tipo de paciente que cada lista representa.

8.3 TESTES

Os discursos foram concedidos pela Ipub (Instituto de psiquiatria da UFRJ), sem autorização para compartilhamento. A princípio, apenas discursos de pacientes saudáveis e de pacientes com esquizofrenia foram coletados.

Os 8 classificadores usados foram: K-means, SVC, Gaussian, Árvore de decisão, Random Forest, MLP, GaussianNB e QDA.

Esses classificadores devem dar um valor entre 0 e 1 para cada frase que analisarem. 0 indica 100% de certeza que o paciente está doente, e 1 indica 100% de certeza que ele está saudável. Esses resultados são depois comparados com os resultados reais, a fim de comparar a acurácia dos classificadores.

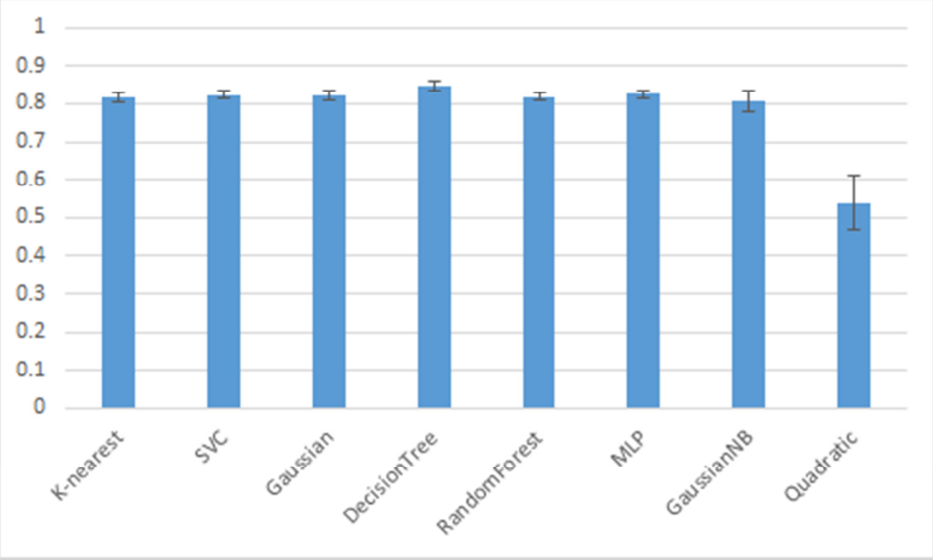
Para análise do desempenho de cada classificador, foi feito um cálculo da acurácia de cada classificador usando seis casas decimais de precisão e um cálculo do intervalo de confiança de cada um.

O cálculo do intervalo de confiança utiliza 3 parâmetros: O tamanho da amostra, o desvio padrão dessa amostra e um valor Alpha, que é um valor entre 0 e 100% que representa a probabilidade de que a média global está de fato no intervalo de confiança. Nesse caso, o tamanho da

amostra foi o número de pacientes (34), o desvio padrão foi calculado usando o resultado da acurácia de 10 treinamentos realizados com cada classificador, e o valor Alpha foi de 90%. e o número de pacientes (34) como parâmetros para a função de confiança de uma distribuição normal. O classificador com a acurácia mais alta foi o DecisionTree, com 0.849905 de acurácia.

Algumas conclusões podem ser obtidas observando esse gráfico. O último classificador (Quadratic) obteve um desempenho muito baixo comparado com os outros. Sua grande margem de erro demonstra a instabilidade do seu resultado, porém mesmo o resultado mais alto dentro dessa margem ainda não é um resultado bom comparado com as outras técnicas. O Quadratic, apesar de lidar bem com variáveis correlacionadas, tenta classificar os dados apenas usando uma função quadrática, o que pode acarretar em um desempenho baixo quando os dados seguem uma distribuição mais complexa ou até mesmo uma distribuição muito simples, como o caso de uma distribuição linear, onde ocorreria overfitting e, consequentemente, uma performance fraca.

Gráfico 8 - Acurácia



Fonte: O autor(2018)

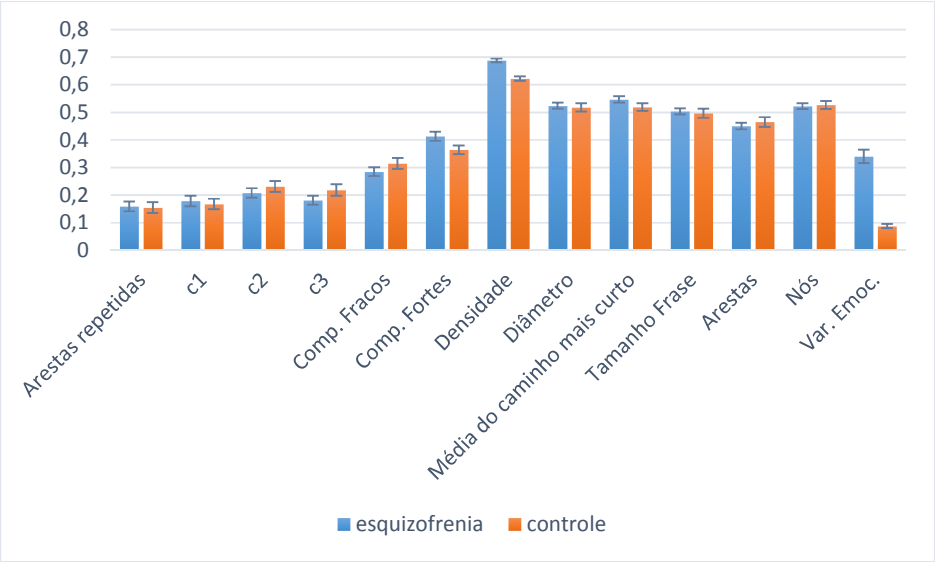
Por outro lado, todos os outros sete classificadores atingiram resultado acima de 0.8, incluindo o SVC, que costuma ter overfitting quando usa dimensões demais para classificar os dados. Isso reflete algo importante sobre a natureza dos dados analisados: São dados que seguem uma distribuição mais complexa do que a quadrática, e que podem ser diferenciados rapidamente usando

alguns parâmetros-chaves, devido ao alto desempenho não só da DecisionTree como da RandomForest. Possivelmente devido a uma quantidade mais alta de dados, estes classificadores obtiveram resultados parecidos, mesmo usando métodos diferentes de classificação.

Algo interessante de se notar foi que o desempenho da DecisionTree foi melhor do que o RandomForest, que usa várias árvores de decisão ligadas por uma feature. Uma analogia que explica bem isso é o ditado popular que diz "Um homem com um relógio sabe que horas são, um homem com dois relógios nunca tem certeza". Uma classificação Random Forest em que uma de suas árvores tenha um desempenho baixo pode comprometer toda a classificação, enquanto que numa Decision Tree, todo o desempenho depende de apenas uma árvore, que será aprimorada durante todo o processo de treinamento.

Outras conclusões importantes foram obtidas quando as médias e os desvios padrão dos dados coletados foram expostos, como no Gráfico 9.

Gráfico 9 – Comparação dos resultados entre os grupos



Fonte: O autor(2018)

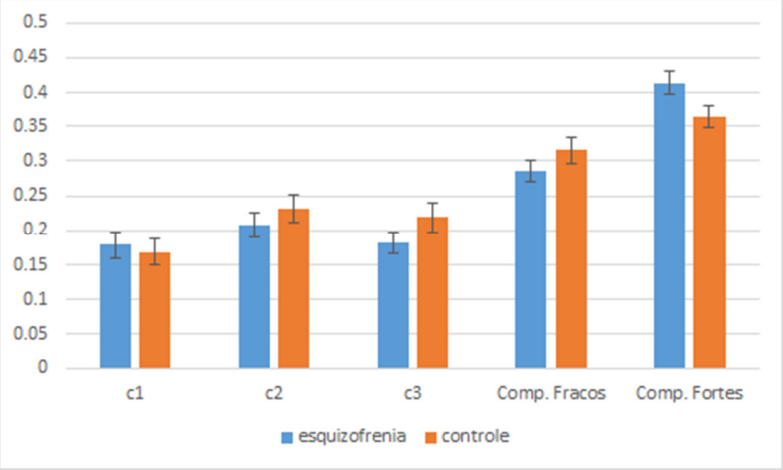
Nesse gráfico, os valores representados são as médias de cada valor apresentado na Seção 8.1, na mesma ordem em que foram exibidos e com um intervalo de confiança, que também utilizou como parâmetros um Alpha de 90% e o número de dados como 23 para o grupo de esquizofrenia e 12 para o grupo de controle, mas nesse caso o desvio padrão foi obtido calculando a raiz quadrada da

variância de cada grupo da base de dados. Os valores foram normalizados para facilitar a visualização dos resultados, o que explica os valores do eixo Y estarem entre 0 e 0.8.

Os pacientes de controle obtiveram um resultado maior em quantidade média de ciclos de qualquer tamanho, quantidade de arestas e de nós, e tiveram um resultado menor em arestas repetidas, densidade do grafo, variação emocional, diâmetro e componentes fortemente conectados. Essa análise será dividida em partes nos Gráficos 10 e 11, que são sub-representações do Gráfico 9.

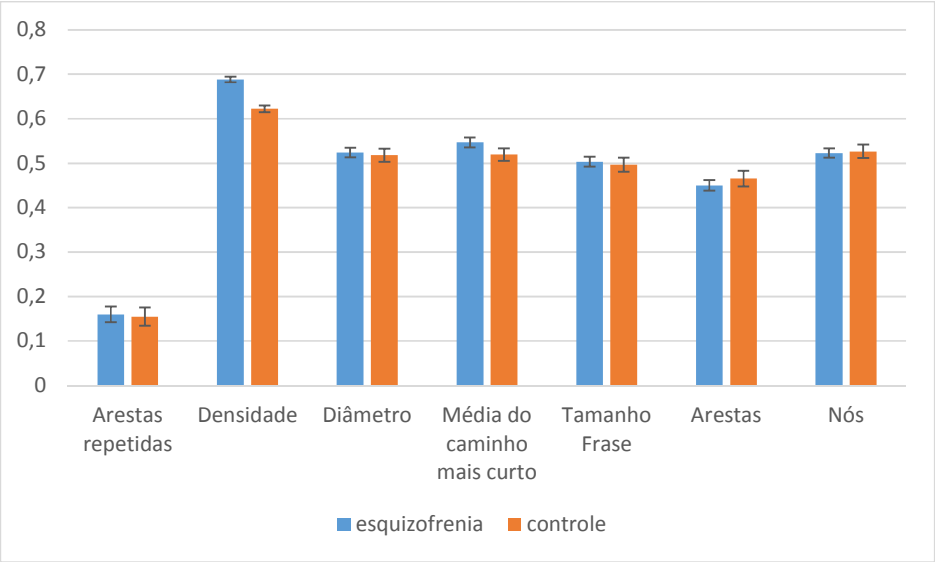
Quanto à análise dos ciclos, nota-se que pacientes de controle obtiveram uma maior quantidade de ciclos de tamanho 2 e 3, mas o tamanho médio do maior ciclo fortemente conectado foi menor do que o obtido por pacientes com esquizofrenia. Essa diferença indica que pacientes sem esquizofrenia possuem mais ciclos em sua fala, porém são ciclos menores, que são sinais de uma fala concisa e foco ao tema. É bastante comum em pacientes com esquizofrenia a fuga ao tema e depois o retorno a ele, e esse aspecto, aliado aos sintomas de logorréia e para-resposta, produzem um grafo pouco conciso e com ciclos maiores.

Gráfico 10 - Comparação de ciclos



Fonte:O autor(2018)

Gráfico 11 - Dados complementares dos grafos



Fonte:O autor(2018)

Quanto aos dados complementares, deve se perceber o tamanho da frase não variou muito entre pacientes de controle e com esquizofrenia. Embora pacientes com esquizofrenia possam desenvolver oligolalia e responder com cada vez menos palavras, pacientes de controle também podem dar respostas igualmente curtas por não sentirem necessidade de fornecer uma resposta mais longa, por exemplo.

A densidade dos grafos de pacientes com esquizofrenia foi relativamente maior comparado com pacientes de controle, embora tenha sido relatado no gráfico anterior que eles possuem uma quantidade menor de ciclos. Pacientes de controle tendem a explicar um assunto de maneira concisa e, quando mudam de assunto, fazem, no seu discurso, uma pequena ligação com o assunto anterior mas rapidamente falam do novo assunto também de maneira concisa. Isso gera um grafo com bastante ciclos, mas com uma densidade geral baixa. Pacientes com esquizofrenia, por não terem a mesma consistência, devido principalmente às suas para-respostas e neologismos, obtêm o efeito contrário, como pode ser visto acima.

A média do caminho mais curto foi ligeiramente maior para pacientes com esquizofrenia, o que normalmente significa uma fraca conexão entre as frases de um parágrafo, típico de um caso de para-resposta ou de neologismo. Normalmente, quanto mais denso um grafo, menor é a distância média entre dois pontos, porém o grafo de um paciente esquizofrênico possui esses dois valores

elevados, o que deve tornar mais fácil reconhecer um grafo desse tipo por um classificador bem treinado.

Analisando o último dado do Gráfico 9, a variação emocional chamou bastante a atenção e pode ter sido um dos principais motivos pelo qual os classificadores obtiveram bons resultados. De fato, uma regressão feita no classificador com melhor desempenho (DecisionTree) revelou que em cerca de 90% dos casos esse classificador conseguia diferenciar um discurso de um paciente com esquizofrenia e um discurso de um paciente de controle analisando apenas a variação emocional da frase. Um paciente saudável consegue, em geral, responder uma pergunta qualquer sem alterar de maneira significativa o sentimento predominante na resposta ou mudar de assunto e fugir do tema da pergunta. Um paciente com esquizofrenia, por outro lado, pode ter diversas mudanças na maneira de se expressar verbalmente, como falar sentenças pouco concisas ou mudar abruptamente o tom de voz, o que se reflete na escolha de palavras usadas. Isso pode tornar a emoção presente em uma frase completamente diferente da emoção presente na frase seguinte, tornando o contexto emocional de um parágrafo bastante instável. Essa diferença foi a mais relevante do gráfico e demonstra que a análise de sentimentos tem um papel muito importante nessa identificação.

De maneira geral, pode-se concluir que os resultados obtidos são consistentes com a literatura atual e que os sintomas da esquizofrenia podem explicar razoavelmente as diferenças métricas entre os dois tipos de pacientes.

TRABALHOS FUTUROS

Um possível desenvolvimento desse trabalho seria o de classificar mais doenças. Para isso, além de uma base de dados que abrangesse qualquer doença adicional, seria necessário um conjunto de classificadores específicos para cada doença. Uma vez que os classificadores utilizados são binários, o uso de um classificador de clusters ou vários classificadores, um para cada doença, poderiam expandir a área de abrangência da classificação.

Classificadores fuzzy também poderiam tornar a classificação mais precisa e fornecer um nível de certeza quanto a cada predição, agindo assim como uma ferramenta de suporte à decisão. Para isso, seria necessário uma medida mais precisa do nível da doença em cada paciente registrado, o que no momento apenas classifica como doente ou saudável, sem um detalhamento do quanto a doença está afetando o paciente.

Quanto à base de dados, uma possível melhora seria ter um conjunto de pacientes com esquizofrenia que ainda não tivessem passado pelo treinamento. Uma grande quantidade desses pacientes, por ter passado pelo treinamento, possuíam menos sintomas da doença, o que torna mais difícil de serem classificados corretamente. Além disso, por não se ter discursos de pacientes antes do treinamento, isso pode tornar os classificadores ineficientes em identificar esses sintomas em pessoas que claramente teriam a doença, e isso pode diminuir a credibilidade da aplicação.

Por fim, dados mais relevantes sobre os pacientes podem melhorar bastante o desempenho dos classificadores. Como foi visto, a simples diferença de idade entre os pacientes do grupo de controle produziu dados bastante diferentes, e o fato de que os classificadores não puderam obter a idade deles pode ter prejudicado o seu desempenho. Outros dados não vistos podem se mostrar relevantes e possivelmente reduzir a quantidade de variáveis analisadas e tornar a classificação mais eficiente.

CONCLUSÃO

Muitas doenças mentais possuem sintomas parecidos e, sem uma identificação adequada, podem não ter o tratamento correto e atrasar a reabilitação do paciente. Muitos pacientes sofrem por receberem um diagnóstico incorreto e muitas vezes têm uma piora no seu quadro clínico.

Além disso, a personalidade de cada paciente pode afetar bastante o seu diagnóstico. Diversos aspectos de uma personalidade são correlacionados com doenças mentais, mas o quanto esses aspectos estão de fato impulsionando uma determinada doença mental ainda é um mistério em alguns casos, como por exemplo o autismo. Por conta disso, um entendimento não só das doenças mentais como também de aspectos importantes da personalidade são necessários para que se possa fazer uma distinção mais precisa e identificar sinais dessas doenças mais cedo.

Análises de discursos de paciente são importantes para a identificação e tratamento de doenças e agem como uma ferramenta de suporte à identificação de doenças mentais. Apesar de existirem diversos métodos para identificar essas doenças, normalmente são usados mais de um desses métodos e ainda assim não se tem 100% de precisão nos resultados. Um método simples e eficiente como a análise de um discurso pode reforçar um resultado correto ou oferecer alternativas para um diagnóstico potencialmente incorreto, auxiliando os médicos e possivelmente salvando vidas.

APÊNDICE A — Perguntas usadas nas entrevistas de pacientes com esquizofrenia

Motivação extrínseca

1. O que te fez querer fazer o treinamento?

2. A ajuda de custo foi o que mais te levou a fazer o treinamento?

3. A ajuda de custo o deixou mais entusiasmado para fazer o treinamento?

4. Você acha que a ajuda da equipe melhora seu desempenho?

5. A equipe ajuda você a se sentir motivado?

6. Qual a sua opinião sobre as exigências do estudo (o tempo de treinamento, a necessidade de freqüência, duração das avaliações, coleta de sangue, etc.)?

7. Você lembra como soube do treinamento?

8. O que você mais gostou na pesquisa? Por quê?

Motivação intrínseca

9. Você gosta/gostou de fazer o treinamento?

10. O que você mais gosta/gostava no treinamento?

11. Você gostaria de saber sobre o seu progresso no treinamento? Acredita que isso iria motivá-lo?

12. Você está empolgado em fazer o treinamento?

13. O treinamento te deixou mais confiante?

14. Você acredita que conseguiria fazer o treinamento sem ajuda?

Desempenho

15. Você acredita que dá/deu seu melhor no treinamento?

16. Você está satisfeito com o seu desempenho no treinamento?

17. Ao fim do treinamento, você ficou contente com o seu desempenho no certificado?

18. Você acredita ter tido uma boa frequência?

Crença

19. Acredita que sua cognição vai melhorar/melhorou por causa do treinamento?

20. O que te deixa/deixou com mais vontade continuar com o treinamento?

ANEXO A — Algoritmo de Johnson

```
from collections import defaultdict
def simple_cycles(G):
    # Yield every elementary cycle in python graph G exactly once
    # Expects a dictionary mapping from vertices to iterables of vertices
    def _unblock(thisnode, blocked, B):
        stack = set([thisnode])
        while stack:
            node = stack.pop()
            if node in blocked:
                blocked.remove(node)
                stack.update(B[node])
                B[node].clear()
    G = {v: set(nbrs) for (v,nbrs) in G.items()} # make a copy of the graph
    sccs = strongly_connected_components(G)
    while sccs:
        scc = sccs.pop()
        startnode = scc.pop()
        path=[startnode]
        blocked = set()
        closed = set()
        blocked.add(startnode)
        B = defaultdict(set)
        stack = [ (startnode,list(G[startnode])) ]
        while stack:
            thisnode, nbrs = stack[-1]
            if nbrs:
                nextnode = nbrs.pop()
                if nextnode == startnode:
                    yield path[:]
```

```
                closed.update(path)
            elif nextnode not in blocked:
                path.append(nextnode)
                stack.append( (nextnode,list(G[nextnode])) )
                closed.discard(nextnode)
                blocked.add(nextnode)
                continue
        if not nbrs:
            if thisnode in closed:
                _unblock(thisnode,blocked,B)
            else:
                for nbr in G[thisnode]:
                    if thisnode not in B[nbr]:
                        B[nbr].add(thisnode)
            stack.pop()
            path.pop()
    remove_node(G, startnode)
    H = subgraph(G, set(scc))
    sccs.extend(strongly_connected_components(H))
def strongly_connected_components(graph):
    # Tarjan's algorithm for finding SCC's
     # Robert Tarjan. "Depth-first search and linear graph algorithms." SIAM journal on computing. 1972.
    # Code by Dries Verdegem, November 2012
    # Downloaded from http://www.logarithmic.net/pfh/blog/01208083168
    index_counter = [0]
    stack = []
    lowlink = {}
    index = {}
    result = []
    def _strong_connect(node):
        index[node] = index_counter[0]
```

```
        lowlink[node] = index_counter[0]
        index_counter[0] += 1
        stack.append(node)
        successors = graph[node]
        for successor in successors:
            if successor not in index:
                _strong_connect(successor)
                lowlink[node] = min(lowlink[node],lowlink[successor])
            elif successor in stack:
                lowlink[node] = min(lowlink[node],index[successor])
        if lowlink[node] == index[node]:
            connected_component = []
            while True:
                successor = stack.pop()
                connected_component.append(successor)
                if successor == node: break
            result.append(connected_component[:])
    for node in graph:
        if node not in index:
            _strong_connect(node)
    return result
def remove_node(G, target):
    # Completely remove a node from the graph
    # Expects values of G to be sets
    del G[target]
    for nbrs in G.values():
        nbrs.discard(target)
def subgraph(G, vertices):
    # Get the subgraph of G induced by set vertices
    # Expects values of G to be sets
    return {v: G[v] & vertices for v in vertices}
##example:
```

```
#graph = {0: [7, 3, 5], 1: [2], 2: [7, 1], 3: [0, 5], 4: [6, 8], 5: [0, 3, 7], 6: [4, 8], 7: [0, 2, 5, 8], 8: [4, 6, 7]}
#print(tuple(simple_cycles(graph)))
```

REFERÊNCIAS

CAD. Historia del deep blue. **Computación Aplicada al Desarrollo S.A. de C.V.** .revista de computación, 2017. Disponível em:<http://www.cad.com.mx/historia_de_deep_blue.htm. >. Acesso em: 15 mai. 2018.

DENECK, Kerstin. **Are sentiwordnet scores suited for multi-domain sentiment classification?** . Proceedings of the fourth international conference on digital information management, p. 1-6, 2009.

POOLE, David et al. **Computational Intelligence: A logical approach.** 1 ed. Oxford university press, New York. 1997.

DENECKE, Kerstin; DENG, Yihan. Sentiment analysis in medical settings: New opportunities and challenges. **Elsevier**. Semmelweisstr, 25 mar. 2015.

E. DELISI, Lynn. Speech disorder in schizophrenia: Review of the literature and exploration of its relation to the uniquely human capacity of language. **Schizophrenia Bulletin** . 2001. 481-496.

FERNANDA, Ferrairo ; BRUNO, Garattoni. **Médicos cometem 12 milhões de erros por ano**. 2016 Disponível em: <https://super.abril.com.br/saude/medicos-cometem-12-milhoes-de-erros-por-ano-so-nos-eua/.> Acesso em: 15 mai. 2018.

NATIONAL INSTITUTE OF HEALTH. Schizophrenia. **NIH**. 1997. Disponível em:<https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml>. Acesso em: 15 mai. 2018.

NUNES, Edeleon Marcelo. **Mineração de textos** : Detecção automática de sentimentos em comentários nas mídias sociais. Belo Horizonte, 2016 Dissertação (Sistema de Informação e Gestão do conhecimento) - UNIVERSIDADE FUMEC, 2016.

RIAZ, Ferozuddin; M. ALI, Khidir. Applications of graph theory in computer science. **Jubail University College**. 2011

HAGBERG, Aric. **NetworkX reference.** 2018. Disponível em: <https://networkx.github.io/documentation/stable/_downloads/networkx_reference.pdf.> Acesso em: 7 jun. 2018.

BERTOLA, Laiss et al. **Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls.** Belo Horizonte. Universidade Federal de Minas Gerais, 2014.

Mota, M. B., Furtado, R., Maia, P. P. C., Copelli, M., Ribeiro, S.. Graph analysis of dream reports is especially informative about psychosis. **Nature.** 15 jan 2014.