Fabrício Velôso de Jesus

Análise e classificação de comentários

Brasil

2018, v1.0

Fabrício Velôso de Jesus

Análise e classificação de comentários

Trabalho monografico apresentado para obtenção do grau de bacharel em ciências exatas e tecnológicas.

Universidade Federal do Recôncavo da Bahia - UFRB Bacharelado em Ciências Exatas e Tecnológicas

Orientador: Tiago Palma Pagano

Brasil

2018, v1.0

Resumo

Palavras-chave:

Abstract

Keywords:

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

BMU	*Best Match Unit*
IA	Sigla para Inteligência Artificial
KDD	*Knowledge Discovery in Database*, em portugês Descoberta de Conhecimento em Bases de Dados
SOM	*Self-Organizing Map*, em portugês Mapas auto organizáveis
RNAs	Sigla para Redes Neurais Artificiais

Lista de símbolos

Γ	Letra grega Gama

Sumário

1 Introdução

1.1 Objetivo

Analisar e classificar comentários de twitter segundo seu caráter misógino.

1.2 Objetivos específicos

Utilizar métodos capazes de classificar os comentários segundo seu caráter misógino. Dentro deste comportamento de aversão às mulheres existem subcategorias, que devem ser declaradas e evidenciadas na classificação.

Analisar caracteristicas comuns as frases que pertecem ao mesmo grupo e determinar a ocorrência e relevância de determinadas palavras para a identificação.

Determinar se tal comportamento possui direcionamento a um usuário em específico, ou é realizado de forma a generalizar todas as mulheres.

1.3 Justificativa

Como consequência, a análise dos resultados obtidos neste trabalho poderá prover um padrão especifico referente ao comportamento de usuários misóginos no twitter.

1.4 Metodologia

Aplicar métodos de mineração de dados em textos para realizar o ajuste dos dados existentes na base.

Utilizar aprendizado de máquina nos dados ajustados para criar uma rotina de classificação das frases. A proposta aqui é com o auxílio de redes neurais, evidenciar dados específicos encontrados em comentários que refletem um cunho misógino, no qual destacamos o método de mapas auto organizáveis com o intuito de evidenciar características comuns em frases que possuem a mesma classificação.

1.5 Problematização

Com auxílio de métodos inerentes a inteligência artificial é possível determinar a existência de misoginia em um comentário?

Através do agrupamento de características é praticável a classificação das frases misóginas em subcategorias?

Existe um padrão para comentários que apresentam cunho misógino?

2 Referêncial Teórico

Neste capítulo as referências conceituais e conceitos envolvidos neste trabalho serão descritos. Partindo da definição de misoginia, passando pelas técnicas envolvidas, e arrematando com as concepções de analise dos dados.

2.1 Misoginia

Segundo Bloch (1995) qualquer definição essencialista da mulher, seja ela negativa ou positiva, feita por um homem ou uma mulher, é a definição fundamental de misoginia. Deodato (2017) também define misoginia como o termo utilizado para caracterizar a antipatia, o desprezo ou a aversão as mulheres.

O constante conflito de ideologias entre gerações mostra como o preconceito surge, tendo em vista as mudanças em âmbitos sociais e em relações de convivência, econômicas, pessoais, entre outras, ao passar dos anos, além da dificuldade de adaptação de costumes e pensamentos. A misoginia, base de vários outros preconceitos é a mais alarmante e evidente entre as discriminações que assolam o Brasil. (MATA; SOARES, 2017)

Conforme Pazó e Júnior (2016) e Álvares (2017) a misoginia em redes sociais é um comportamento brutal e insultuoso, que vão desde feedback negativo à aparência a ameaças mais sérias. Onde os perseguidores se valem da principal característica da navegação, o anonimato, para disseminarem seus ataques.

A misoginia pode se manifestar de diferentes maneiras e manifestações ofensivas às mulheres e não é raro de se encontrar em diversos ambientes virtuais. Não apenas nas mais utilizadas redes sociais virtuais, como também em páginas voltadas para discussões anônimas.(PAZÓ; JÚNIOR, 2016)

2.1.1 Classificação de comportamento misógino

O comportamento misógino em um tweet deve ser classificado como pertencente a uma das seguintes categorias, conforme descrito por (FERSINI; NOZZA; ROSSO, 2018):

- *Stereotype & Objectitication* (Estereótipo & Objetificação): uma idéia ou imagem de mulher amplamente difundida, porém fixa e simplista; descrição do físico feminino, apelo e/ou comparações a padrões delimitados;
- *Dominance* (Domínio): afirmar a superioridade de homens sobre as mulheres para destacar a desigualdade de gênero;

- *Derailing*: justificar o abuso de mulheres, rejeitando a responsabilidade masculina; uma tentativa de interronper a conversa, a fim de redirecionar as conversas das mulheres em algo mais confortável para os homens;

- *Sexual Harassment & Threats of Violence*(Assédio sexual & Ameaças de violência): descrever ações como avanços sexuais, pedidos de favores sexuais, assédio de natureza sexual; intenção de afirmar fisicamente poder sobre as mulhers através de ameaças de violência;

- *Discredit* (Discrédito): insultar as mulheres sem nenhuma intenção maior.

2.2 Mineração de Dados

Conforme Amorim (2006) e Santos (2008) a definição de mineração de dados (*Data Mining*) pode ser descrita como o conjunto de técnicas que permite a extração de conhecimentos, padrões e relações de grandes massas de dados que não seriam descobertas com facilidade a olho nu pelo homem.

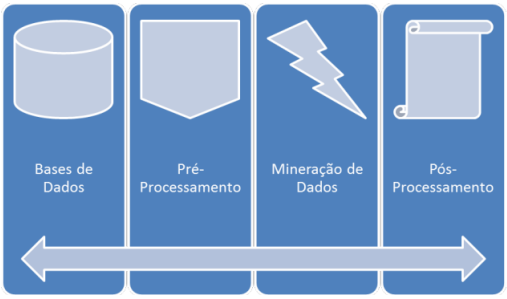
A descoberta de padrões constitui-se de um processo que se inicia pela escolha dos dados que documentam de alguma maneira a pergunta que o especialista deseja responder. Os dados são integrados e pré-processados para que sejam entregues estruturados, higienizados, selecionados e padronizados à tarefa de mineração de dados. Na tarefa de mineração aplica-se alguma técnica inteligente capaz de encontrar soluções que auxiliam o especialista na descoberta de uma resposta. O resultado desta tarefa deve ser pós-processado para que se apresentem análises qualitativas e/ou quantitativa dos elementos encontrados e, quando possível, apresentados de maneira que possa ser interpretado de maneira a facilitar a tomada de decisão. (SILVA; SILVA, 2014, p.569 - p.570)

Todo este processo citado acima, pode ser chamado de Descoberta de Conhecimento em Base de Dados ou simplismente KDD, consoante Fayyad, Piatetsky-Shapiro e Smyth (1996 apud SILVA; SILVA, 2014).

Uma das definições mais utilizadas para o termo KDD é a de Fayyad, que o define como "um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis".(CAMILO; SILVA, 2009, p.3)

Ainda segundo Camilo e Silva (2009), até agora não é consenso a definição dos termos *Data Mining* e *KDD*. No entanto, todos concordam que o processo de mineração deve ser interativo, iterativo e particionado em fases, comforme visto na Figura 1.

Figura 1 – Processo de Descoberta de Conhecimento em Base de Dados



Fonte: Silva e Silva (2014, p.570)

2.2.1 Tipos de estrutura de dados

Segundo Sargiani et al. (2018) a estrutura com a qual os dados são apresentados é importante, ela induz de forma direta nas ferramentas que serão utilizadas e nas técnicas de tratamento. Na análise de dados não estruturados ferramentas que permitem extração de conhecimento a partir de dados sem estrutura são utilizadas. Para dados semi-estruturados as técnicas são definidas com base no caso em específico. E para dados estruturados bancos relacionais são utilizados.

De acordo com Morais e Ambrósio (2007) e Sargiani et al. (2018), o processo de descoberta de conhecimento em **dados estruturados** é feito através do uso de ferramentas baseadas em métodos estatísticos, métodos provenientes da área de recuperação de informações e ontologias, estes dados possuem um formato definido através de algum critério. A informação pode ser representada em *datasets*, tabelas, arquivos multimídia ou arquivos texto.

Os **dados semi-estruturados** não possuem um formato adequado para o uso de apenas uma ontologia, porém podem ser identificados, pois possuem algum grau de regularidade.(BARROS et al., 2008 apud SARGIANI et al., 2018)

Conforme Cecilio e Castro (2015) e Sargiani et al. (2018) os chamados **dados não estruturados** ocorrem quando a informação não possui nenhum formato reconhecível, comumente correlacionado a linguagem natural, o Twitter, e-mails e conteúdo em fóruns são alguns exemplos, as informações não estão dispostas em tabelas númericas organizadas em linhas e colunas e não possuem um formato adequado para o uso de uma ontologia. A mineração de textos é uma técnica usada pelos sistemas inteligentes que engloba o processamento de dados não estruturados do tipo texto, em outras palavras as *strings* ou sequência de caracteres de um texto.

2.3 Mineração de Textos

A mineração de textos consiste em extrair regularidades , padrões ou tendências para determinados objetivos, essa obtenção de informação é feita em grandes volumes de textos em linguagem natural. É um campo novo e multidisciplinar que inclui conhecimento de áreas como Estatística, Informática, Linguistica e Ciência Cognitiva. Aranha e Passos (2006)

Conforme Sargiani et al. (2018) para a execução da análise é necessário estruturar esse tipo de dados não estruturados por meio de um modelo de representação, que transforma os termos de cada publicação em um valor de relevância. A análise de dados não estruturados é feita em três fases distintas.

- A primeira fase do processo é a construção do vacábulário que se dá pelo processo de mineração de textos. Isto é, o texto com todos os comentários selecionados representa o *corpus* inicial. Para a geração da representação (*corpus* representado) é necessário seguir os seguintes passos, como descrito por Goker e Davies (2009 apud SARGIANI et al., 2018) e Silva, Peres e Boscarioli (2017 apud SARGIANI et al., 2018):

    1. *Tokenization*: A partir do caractere espaço, os comandos das instruções são separados em tokens. Os caracteres especiais como vírgulas (","), e pontuação em geral, são removidos, assim como números. Padronização de capitalização para minúsculas(ou maiúsculas) também é feita nesta fase;

    2. *Stopwords*: Palavras como artigos, advérbios, pronomes, preposições, que são comuns em diferentes contextos, são removidos do processo;

    3. *Stemming*: As palavras resultantes das etapas anteriores passam por uma normalização ortográfica para que sejam reduzidas ao radical. Este processo é importante, pois permite que palavras com o mesmo radical sejam consideradas como semelhantes.

- A segunda fase é a geração do *corpus*. Este *corpus* é uma matriz contendo todos os documentos analisados, todos os termos encontrados, e suas respectivas quantidades em cada documento;

- A última fase é a geração da matriz de frequências, momento em que é feita a relação entre cada documento e os termos constantes. O formato ideal a ser escolhido depende principalmente da análise que será feita posteriormente, pois o formato dessa matriz afeta diretamente no processo de análise.

2.4 Inteligência Artificial

Segundo Fernandes (2005) a inteligência artificial é a parte da Ciência da Computação voltada para o desenvolvimento de sistemas de computadores inteligentes, isto é, sistemas que exibem características que estão associadas à inteligência no comportamento humano, como compreensão da linguagem, aprendizado, raciocínio, resolução de problemas, entre outros.

De acordo com Hodges (1999) o **teste de Turing**, proposto por Alan Turing(1950), fornece uma definição operacional satisfatória de inteligência. Seu objetivo é descobrir se uma IA é inteligente a ponto de enganar um humano, de forma que ele acredite que uma pessoa está respondendo suas perguntas feitas e respondidas através de textos. O argumento de Turing é simplesmente o de que o cérebro deve também ser considerado uma máquina de estado discreto e que as únicas características do cérebro relevantes para o pensamento ou a inteligência são aquelas situadas no nível de descrição da máquina de estado discreto, portanto a materialização física é irrelevante.

Para que uma IA passe no teste de Turing, ela deve apresentar as seguintes capacidades, como descrito por RUSSELL e NORVIG (2013)

- **processamento de linguagem natural** para permitir que ele se cominique com sucesso em um idioma natural;
- **raciocínio automatizado** para usar as informações armazenadas com a finalidade de responder a perguntas e tirar novas conclusões;
- **representação de conhecimento** para armazenar o que sabe ou ouve;
- **aprendizado de máquina** para se adaptar a novas circunstâncias e para detectar e extrapolar padrões.

Conforme RUSSELL e NORVIG (2013) o primeiro trabalho reconhecido como IA foi proposto por Warren McCulloch e Walter Pitts (1943). Este trabalho foi baseado em três fontes: o conhecimento da fisioloia básica e da função dos neurônios no cérebro; a teoria da computação de Turing; e uma análise formal da lógica proposicional criado por Russell e Whitehead. Esses pesquisadores propuseram um modelo de neurônios artificiais, onde cada neurônio se caracteriza por estar "ligado"ou "desligado", com a troca para "ligado"ocorrendo em resposta à estimulação por um número suficiente de neurônios vizinhos. O estado era considerado "equivalente em termos concretos a uma proposição que definia seu estimulo adequado". Eles mostraram que qualquer função computável podia ser calculada por certa rede de neurônios conectados e que todos os conectivos lógicos podiam ser implementados por estruturas de redes simples.

Vários trabalhos que podem ser caracterizados como IA surgiram, mas a visão proposta por Alan Tuting foi talvez a mais influente. Em 1947, ele proferia palestras sobre o tema na Sociedade Matemática de Londres e articulou um programa de trabalhos persuasivo em seu artigo de 1950, "computing Machinery and Intelligence"Hodges (1999). Artigo no qual apresentou o teste de Turing, algoritmos genéticos, aprendizagem de máquina e aprendizagem por reforço.

Ainda segundo RUSSELL e NORVIG (2013) os pesquisadores da IA possuiam prognósticos ousados de seus sucessos futuros, porém entre 1966 e 1973 alguns tipos de dificuldades surgiram:

1. Primeiro tipo de dificuldade surgiu porque a maioria dos primeiros programas não tinha conhecimento de seu assunto, isto é, eles obtiam sucesso por meio de manipulações sintáticas simples;

2. O segundo tipo de dificuldade foi a impossibilidade de tratar muitos problemas que a IA estava tentando resolver, a maior parte dos primeiros programas de IA resolvia problemas experimentando diferentes combinações de passos até encontrar a solução.*O fato de um programa poder encontrar uma solução em princípio não significa que o programa contenha quaisquer dos mecanismos necessários para encontrá-la na prótica*;

3. Uma terceira dificuldade surgiu devido a algumas limitações fundamentais nas estruturas básicas que estavam sendo utilizadas para gerar a comportamento inteligente.

Os chamados modelos **conexionistas** para sistemas inteligentes eram vistos por alguns como concorrentes diretos dos modelos simbólicos promovidos por Newell e Simon e da abordagem logicista de McCarthy e outros pesquisadores Smolensky (1988 apud RUSSELL; NORVIG, 2013).

Pode parecer óbvio que, em certo nível, os seres humanos manipulam símbolos, mas os conexionistas mais fervorosos questionavam se a manipulação de símbolos tinha qualquer função explicativa real em modelos detalhados de cognição. Essa pergunta permanece sem resposta, mas a visão atual é de que as abordagens conexionista e simbólica são complementares, e não concorrentes. Como ocorreu com a separação da IA e da ciência cognitiva, a pesquisa moderna de rede neural se bifurcou em dois campos, um preocupado com a criação de algoritmos e arquiteturas de rede eficazes e a compreensão de suas propriedades matemáticas, o outro preocupado com a modelagem cuidadosa das propriedades empíricas de neurônios reais e conjuntos de neurônios.(RUSSELL; NORVIG, 2013)

2.5 Redes Neurais Artificiais

Segundo Braga, Carvalho e Ludermir (2000) RNAs são sistemas paralelos distribuidos compostos por unidades de processamento simples (nodos) que calculam determinadas funções matemáticas (normalmente não-lineares). Essa unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais. Estes modelos de conexões normalmente estão associados a pesos, os quais aramazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede. O funcionamento destas redes é inspirado em uma estrutura física natural: o cérebro humano.

A abordagem conexionista ficou adormecida durante os anos 70, porém alguns pesquisadores continuaram desenvolvendo trabalhos na área. Dentre eles podem ser citados Igor Aleksander (redes sem pesos) na Inglaterra, Kunihiko Fukushima (cognitron e neocognitron) no Japão, Steven Grossberg (sistemas auto-adaptativos) nos EUA, e Teuvo Kohonen (memórias associativas e auto-organizadas) na Finlândia.

2.5.1 Motivação para as RNAs: redes biológicas

O cérebro humano é um imenso e complexo bosque de células e conexões intercelulares. Esse bosque emaranhado é composto de aproximadamente 100 bilhões de neurônios $(1 * 10^{11})$ de formas e tamanhos diferentes. Considera-se que apenas no córtex cerebral, que contém quase a metade desse número, isto é, cerca de 50 bilhões, existam mais de 500 tipos de neurônios morfologicamente diferentes, distribuídos em 52 áreas.(MORA, 2016, p.18)

A estrutura dos nodos, a topologia dessas conexões e o comportamento conjunto dos neurônios naturais constroem a base de estudo das RNAs. As RNAs tendem a reproduzir as funções das redes biológicas, buscando colocar em prática a sua dinâmica e seu comportamento básico.

Conforme Braga, Carvalho e Ludermir (2000), como caracteristicas comuns, ambos os sistemas são baseados em unidades de computação paralela e distribuída que se comunicam por meio de conexões sinápticas, possuem detetores de características, redundância e modularização das conexões. Apesar de pouca similaridade entre os dois sistemas do ponto de vista biológico, estas características semelhantes permitem às RNAs reproduzirem com fidelidade várias funções inerentes dos seres humanos

2.6 Perceptrons Multicamadas

Os *perceptrons* multicamadas ou MLPs se caracterizam pela presença de uma ou mais camadas itermediárias ou escondidas(camadas em que os neurônios são efetivamente

unidades processadoras, mas não correspondem à camada de saída). Adicionando-se uma ou mais camadas itermediárias, aumenta-se o poder computacional de processamento não-linear e armazenagem da rede. Em uma única camada oculta, suficientemente grande, é possivel representar, com exatidão qualquer função continua das entradas. O conjunto de saídas dos neurônios de cada camada da rede é utilizada como entrada para a camada seguinte.(DUARTE et al., 2009, p.40)

Comforme Braga, Carvalho e Ludermir (2000) e Kovács (2002), por volta do fim da década de 1950, na Universidade de Cornell, Rosenblatt deu continuidade às idéias de McCulloch. Criando uma genuína rede de múltiplos neurônios do tipo *discriminadores lineares* esta foi o seu novo modelo, o *perceptron*. Um perceptron é uma rede com a seguinte topologia, os neurônios são dispostos em três *camadas*. A primeira recebe as entradas do exterior e possui conexões fixas (retina); a segunda recebe a saída da camada de entrada, através de conexões cuja eficiência de transmissão (peso) é ajustável constituem a segunda camada e, por sua vez, envia saídas para a ultima camada (resposta).

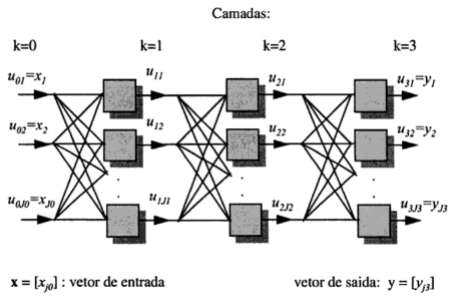
Ainda segundo Kovács (2002), Braga, Carvalho e Ludermir (2000) e RUSSELL e NORVIG (2013) o problema que Rosenblatt propôs a resolver foi o de casos simples com implementação de funções booleanas **E** e **OU** de duas variáveis, que são problemas linearmente separáveis, isto é, problemas cuja solução pode ser obtida ao dividir o espaço de entrada em duas regiões através de uma reta. O perceptron, não consegue detectar conectividade, paridade e simetria, que são problemas não-linearmente separáveis. Estes são exemplos de *hard learning problems* (problemas difíceis de aprender). Porém qualquer funcionalidade desejada pode ser obtida ligando um grande número de unidades em redes de profundidade arbitrária.

Com referência à Figura 2.Uma rede neural multicamada de K camadas, terá como entrada um vetor $\mathbf{x}$ de dimensão J_0 de componentes $x_{j_0}, j_0 = 1, 2, ...J_0$. Estas conectam-se às entradas dos J_1 neurônios numa primeira camada. As saídas $u_l j_1, j_1 = 1, 2, ...J_1$ destes, formando as componentes de um novo vetor $\mathbf{u}_1$ de dimensão J_1, conectam-se às entradas dos J_2 neurônios da camada seguinte e assim sucessivamente até a última camada que consistirá de J_K neurônios fornecendo como saída da rede um vetor $\mathbf{y} = \mathbf{u}_K$ de dimensão J_K. Genéricamente, u_{kj_k} denota a saída do j_k -ésima entrada da rede, e para $k = K$ a j_k -ésima saída da rede.(KOVÁCS, 2002, p. 39–40)

2.6.1 Aprendizagem

Segundo Braga, Carvalho e Ludermir (2000) existem vários algoritimos para treinar redes MLP, porém o mais conhecido é o *back-propagation*. A maioria dos métodos de aprendizado para RNAs do tipo MLP utiliza variações desse algoritmo. O algoritmo *back-propagation* é um algoritmo supervisionado que utiliza pares de entrada e saida desejada para ajustar os pesos da rede por meio de um mecanismo de correção de erros.

Figura 2 – Rede Neural Multicamada



Fonte: Kovács (2002, p.40)

Comforme RUSSELL e NORVIG (2013) as redes neurais são capazes de tarefas muito complexas de aprendizagem, porém é necessário certa qantidade de esforço para obter a estrutura correta de rede para que a convergência para algo próximo ao ótimo global no espaço de peso seja alcançada.

O treinamento do algoritmo de *back-propagation* ocorre em duas etapas que percorrem a rede em sentidos opostos, o *forward* define a saida da rede para uma determinado padrão de entrada e o *backward* utiliza a saída desejada e a saída fornecida pela rede para atualizar os pesos de suas conexões, o algoritmo segue os seguites passos conforme explanado por Braga, Carvalho e Ludermir (2000):

1. Inicializar pesos e parâmetros;

2. Repetir até o erro ser mínimo ou até a realização de um dado número de ciclos:

   2.1. Para cada padrão de treinamento X;

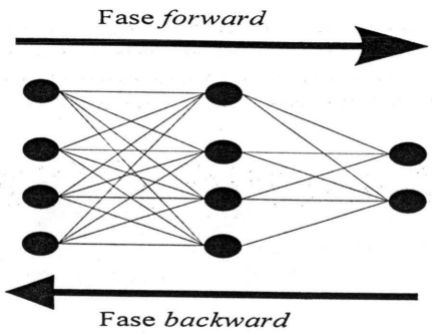
      2.1.1. Definir saída da rede através da fase *forward*;

      2.1.2. Comparar saídas produzidas com as saídas desejadas;

      2.1.3. atualizar pesos do nodos através da fase *backward*.

Ainda segundo Braga, Carvalho e Ludermir (2000) este algoritmo propõe uma forma de definir o erro dos nodos das camadas intermediárias, possibilitando o ajuste de seus pesos, e no seu fluxo de processamento os dados seguem da entrada para a saída no sentido *forward*, e os erros, da saída para a entrada no sentido *backward*, conforme descrito na Figura 3 .

Figura 3 – Fluxo de processamento do algoritmo *back-propagation*



Fonte: Braga, Carvalho e Ludermir (2000, p.60)

2.7 Mapas Auto Organizáveis de Kohonen

Kohonen (2013) define SOM como uma técnica de análise de dados não supervisionado, realizada por meio do treinamento de um *grid* de neurônios. A anélise de sinais chamada *"Vector Quantization"*, introduzido por Forgy (1965) para forma vetorial, e por Lloyd (1982) para dados escalares, foi a base para a criação desta técnica.

O SOM realiza uma projeção não linear do espaço de dados de entrada, em R^D, para o espaço de dados do arranjo, em R^P, executando uma redução dimensional quando $P < D$. Como o arranjo é normalmente unidimensional ou bidimensional, então $P = 1$ ou $P = 2$. Ao realizar esta projeção não linear, o algoritmo tenta preservar ao máximo a topologia do espaço original, ou seja, procura fazer com que neurônios de vizinhos no arranjo apresentem vetores de pesos que retratem as relações de vizinhança entre os dados. Para tanto, os neurônios competem para representar cada dado, e o neurônio vencedor tem seu vetor de pesos ajustados na direção do dado. Esta redução da dimensionalidade com preservação topologica permite ampliar a capacidade de análise de agrupamentos dos dados pertecentes a espaços de elevada dimensão.(ZUCHINI et al., 2003, p.37)

2.7.1 Definição Matemática do SOM

Kohonen e Honkela (2007) faz a seguinte definição matématica para o SOM:

Os primeiros itens de dados que são vetores euclidianos de n-dimensões devem ser considerados

$$x(t) = [\xi_1(t), \xi_2(t) \ldots, \xi_n(t)].$$

Para uma determinada sequência, t representa o índice de cada item. Onde o i-ésimo modelo será $m_i(t) = [\mu_{i1}(t), \mu_{i2}(t), \ldots, \mu_{in}(t)]$, desta vez t representa o indice na sequência em que são gerados os modelos. Essa sequência é um processo de suavização no qual o novo valor $m_i(t+1)$ é computado iterativamente a contar do valor antigo $m_i(t)$ e o novo

item $x(t)$, determinado por

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci(t)}[x(t) - m_i(t)].$$

Onde $\alpha(t)$ é um fator escalar que define o tamanho da correção, esse valor decresce a cada época, isto é, a cada novo indice t. O indice i se refere ao modelo em processamento, enquanto o indice c refere-se ao modelo com a menor distanfica de $x(t)$ no espaço euclidiano. O fator $h_{ci}(t)$ é a função de vizinhança. É igual a 1 quando $i = c$ e o valor decresce a medida que a distância entre os modelos m_i e m_C aumenta no grid. Além diso, a cada época a largura espacial do kernel no grid diminui. As funções que determinam a convergência, devem ser escolhidas com muita cautela.

Ainda segundo Kohonen e Honkela (2007) embora o algoritmo iterativo tenha sido usado com êxito em incontáveis aplicações, foi expôsto que o esquema *Batch Map* produz resultados essencialmente semelhantes, mas com maior velocidade. A idéia básica é que para cada no j no grid, a média $\bar{x}(t)$ de todos os itens de entrada $x(t)$ é formada e tem m_j como modelo mais próximo. Após isso, os novos modelos são computados como

$$m_i = \sum_j n_j h_{ji} \bar{x}_j / \sum_j n_j h_{ji}$$

onde n_j é o número de itens de entrada mapeados para o nó j, e o índice j é executado para toda vizinhançado do nó i. Esse esquema é iterado alguma vezes para o m_i atualizado, sempre utilizando o mesmo lote de itens de dados de entrada para determinar o novo $\bar{x}_j$.

2.7.2 Aprendizagem

O algoritmo básico de trinamento do SOM consiste em três fases. Na primeira fase, competitiva, os neurônios da camada de saída competem entre si, segundo algum critério, geralmente a distância Euclideana, para encontrar um único vencedor, também chamado de BMU (*Best Match Unit*). Na segunda fase, cooperativa, é definida a vizinhança deste neurônio. Na ultima fase, adaptativa, os vetores de peso do neurônio vencedor e de sua vizinhança são ajustados.(SILVA, 2004, p.34)

Conforme Sassi (2006) um mapa entre um espaço discreto de M neurônios e um espaço N-dimensional contínuo de vetores x de entrada, é implementada em uma rede do tipo o-vencedor-fica-com-tudo. Neste caso um neurônio vencedor pode representar mais de um vetor x. Portanto, este neurônio será o representante de um grupo de padrões x que o fazem ser vencedor. Esse neurônio é adaptado (*winner-takes-all*). O vetor de pesos associado a ele é atualizado de forma a representar ainda mais o dado apresentado, aumentando a probabilidde de que este mesmo neurônio volte a vencer em uma próxima apresentação de dado semelhante, ou do mesmo dado.

Segundo Zuchini et al. (2003) e Sassi (2006) a idéia fundamental na fase cooperativa e adaptiva, é que neurônios próximos no arranjo representem dados próximos no espaço

de dados. A introdução da função de vizinhança faz com que o vetor de pesos não apenas do neurônio vencedor seja atualizado na direção da entrada atual, mas o vetor de pesos dos neurônios que fazem parte de sua vizinhança. Sassi (2006) faz a seguinte analogia, todos o vetores de pesos dos neurônios da rede SOM estariam ligados por elásticos, onde os elásticos de maior intesidade estariam unindo os vetores mais proximos, e conforme a distância entre esses vetores aumenta a intensidade dos elásticos diminui. Quando o vetor peso de um determinado neurônio vencedor fosse alterado, ele arrastaria consigo os demais vetores que estão ligados a ele, e mais intensamente aqueles mais próximos.

2.8 K-means

Conforme Lopes (2004) *clustering* tem sido empregado em um número de aplicações baseadas em texto tais como Recuperação de informação e Categorização de textos. O *clustering* é o agrupamento de representaçãoes similares de documentos em separações onde os documenos pertencentes a mesma partição possuem uma similaridade maior entre eles do que a qualquer outro documento em qualquer outra partição.

O algoritmo K-means, criado por MacQueen em 1967 é o algoritmo de *clustering* mais conhecido e utilizado já que é de aplicação muito simples e eficaz. Segue um procedimento simples de classificação de um conjunto de objetos em um determinado número K de *clusters*, onde o K é determinado a priori.(CAMBRONERO; MORENO, 2006)

Ainda segundo Cambronero e Moreno (2006) o algoritmo recebe tal nome por causa do seu funcionamento, já que o mesmo cria *clusters* a partir da média (ou média ponderada) de seus pontos, e assim estabelece o centróide. Este tipo de representação tem a vantagem de possuir um significado gráfico e estátistico imediato. Cada agrupamento é caracterizado pelo seu centróide, que por sua vez está no centro e na média de todos os elementos que compõem o *cluster*.

Dado um número fixo de k, o *clustering* K-means cria um conjunto de k *clusters* e distribui o conjunto de documentos dados entre esses *clusters* usando a similaridade entre os vetores-documento e os centróides dos *clusters*, Um centróide é o vetor médio de todos os vetores-documento no respectivo *cluster*. Cada vez que se adiciona um documento em um *cluster*, o centróide daquele *cluster* é recalculado. Note que quase sempre, um centróide não corresponde a um documento. A similaridade entre um documento d e um centróide c é calculada como o somatório de todos os vetores-documento no *cluster* dividido pelo número de vetores-documento.(LOPES, 2004, p.53)

2.9 Visualização de dados

De acordo com Freitas et al. (2008) a aplicação do termo visualização hoje é associada à possibilidade de explorar as informaçãoes subjacentes à representação gráfica. Porém genericamente, visualização é entendida como a "representação grafica de dados ou conceitos"Ware (2001 apud FREITAS et al., 2008).

Conforme descrito por Nascimento e Ferreira (2005) o processo de visualização está associado a geração de imagens (mentais ou reais) que possam ser analisadas pelos seres humanos a partir de algo abstrato. O objetivo final é facilitar o entendimento de um assunto determinado, onde o mesmo exigiria maior esforço para ser compreendido. Uma vez que técnicas que facilitam o entendimento de informações a partir de representações visuais de dados são agregados a área de Visualização de Informações, o estudo neste campo apresenta grande utilidade. Na área da Computação, a visualização de informações possue um destaque especial nas áreas de mineração de dados e na engenharia de software, pois auxilia na análise e no entendimento de determinadas estruturas com um maior nivel de abstração.

2.9.1 Word cloud

Conforme Sargiani et al. (2018) nuvens de palavras (*word cloud*) são visualizações gráficas onde as palavras, de acordo com sua relevância dentro do *Corpus* de origem, assumem posição e tamanho defirentes. Cores diferentes também podem ser usadas para auxiliar na diferenciação das palavras, e a própria nuvem de palavras pode ser formatada de acordo com o objetivo de transmissão de conhecimento que se espera.

Como exposto por Heimerl et al. (2014) uma popular área de aplicação para *word cloud* é a sumarização de texto. Neste contexto as nuvens são utilizadas para que uma visão geral intuitiva e visualmente atraente seja fornecida. Tal sumarização é proveitosa para aprendizagem sobre o número e tipo de temas presentes no corpo do texto. Essa visão estatística é gerada pelo correlacionamento entre o tamanho da fonte das palavras representadas com a frequência das mesmas.

Esta técnica foi originada on-line na década de 1990 como nuvens de tags, que foram utilizadas para exibir a popularidade de palavras-chave em bookmarks, consoante Harris (2011).

3 Desenvolvimento

3.1 Obtenção dos Dados

3.2 Pré Processamento

3.3 Análise

3.4 Treinamento SOM e K-Means

3.5 Treinamento MLP

4 Testes e Análise de Resultados

5 Conclusão

Referências

ÁLVARES, C. Pós-feminismo, misoginia online e a despolitização do privado. *Media & Jornalismo*, Centro de Investigação Media e Jornalismo, v. 17, n. 30, p. 99–110, 2017. Citado na página 19.

AMORIM, T. Conceitos, técnicas, ferramentas e aplicações de mineração de dados para gerar conhecimento a partir de bases de dados. *Monografia (Bacharel em Ciência da Computação). Universidade Federal de Pernambuco*, 2006. Citado na página 20.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, v. 5, n. 2, 2006. Citado na página 22.

BARROS, F. A. et al. Hidden markov models and text classifiers for information extraction on semi-structured texts. In: IEEE. *Hybrid Intelligent Systems, 2008. HIS'08. Eighth International Conference on*. [S.l.], 2008. p. 417–422. Citado na página 21.

BLOCH, R. H. *Misoginia medieval*. [S.l.]: Editora 34, 1995. Citado na página 19.

BRAGA, A. d. P.; CARVALHO, A.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: Livros Técnicos e Científicos Rio de Janeiro, 2000. Citado 4 vezes nas páginas 25, 26, 27 e 28.

CAMBRONERO, C. G.; MORENO, I. G. Algoritmos de aprendizaje: knn & kmeans. *Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid*, 2006. Citado na página 30.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, p. 1–29, 2009. Citado na página 20.

CECILIO, S.; CASTRO, R. Análise de dados não estruturados: Mineração de textos. In: _____. [S.l.: s.n.], 2015. p. 79–98. Citado na página 21.

DEODATO, L. *O que é misoginia?* 2017. Disponível em: <https://www.ovalordofeminino.com.br/artigo/o-que-%C3%A9-misoginia>. Citado na página 19.

DUARTE, V. A. R. et al. Mp-draughts-um sistema multiagente de aprendizagem automática para damas baseado em redes neurais de kohonen e perceptron multicamadas. Universidade Federal de Uberlândia, 2009. Citado na página 26.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996. Citado na página 20.

FERNANDES, A. M. da R. *Inteligência Artificial: noções gerais*. [S.l.]: Visual Books, 2005. Citado na página 23.

FERSINI, E.; NOZZA, D.; ROSSO, P. Overview of the evalita 2018 task on automatic misogyny identification (ami). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR. org*, 2018. Citado na página 19.

FORGY, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, v. 21, p. 768–769, 1965. Citado na página 28.

FREITAS, C. et al. Extração de conhecimento e análise visual de redes sociais. *SEMISH-Seminário Integrado de Software e Hardware, Belém do Pará, Brasil, SBC*, p. 106–120, 2008. Citado na página 31.

GOKER, A.; DAVIES, J. *Information retrieval: searching in the 21st century*. [S.l.]: John Wiley & Sons, 2009. Citado na página 22.

HARRIS, J. Word clouds considered harmful. *Nieman Journalism Lab*, 2011. Citado na página 31.

HEIMERL, F. et al. Word cloud explorer: Text analytics based on word clouds. In: IEEE. *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. [S.l.], 2014. p. 1833–1842. Citado na página 31.

HODGES, A. *Turing um filósofo da natureza*. [S.l.]: Unesp, 1999. Citado 2 vezes nas páginas 23 e 24.

KOHONEN, T. Essentials of the self-organizing map. *Neural networks*, Elsevier, v. 37, p. 52–65, 2013. Citado na página 28.

KOHONEN, T.; HONKELA, T. Kohonen network. *Scholarpedia*, v. 2, n. 1, p. 1568, 2007. Citado 2 vezes nas páginas 28 e 29.

KOVÁCS, Z. L. *Redes neurais artificiais*. [S.l.]: Editora Livraria da Fisica, 2002. Citado 2 vezes nas páginas 26 e 27.

LLOYD, S. Least squares quantization in pcm. *IEEE transactions on information theory*, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 28.

LOPES, M. C. S. Mineração de dados textuais utilizando técnicas de clustering para o idioma português. *Rio de Janeiro: sn*, 2004. Citado na página 30.

MATA, E. M. S. da; SOARES, F. H. M. Valores sociais e a construção da misoginia. In: *Congresso Interdisciplinar-ISSN: 2595-7732*. [S.l.: s.n.], 2017. Citado na página 19.

MORA, F. *Continuum: Como Funciona o Cérebro?* [S.l.]: Artmed Editora, 2016. Citado na página 25.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. *Relatório Técnico–Instituto de Informática (UFG)*, 2007. Citado na página 21.

NASCIMENTO, H. A. D.; FERREIRA, C. B. Visualização de informações–uma abordagem prática. In: *XXV Congresso da Sociedade Brasileira de Computação, XXIV JAI. UNISINOS, S. Leopoldo–RS*. [S.l.: s.n.], 2005. Citado na página 31.

PAZÓ, C. G.; JÚNIOR, R. F. M. Misoginia, internet y punitivismo–la investigación de una solución adecuada1. 2016. Citado na página 19.

RUSSELL, S.; NORVIG, P. *Inteligência Artificial. Tradução da Terceira Edição*. [S.l.]: Editora Elsevier, 2013. Citado 4 vezes nas páginas 23, 24, 26 e 27.

SANTOS, R. Computação e matemática aplicada às ciências e tecnologias espaciais, chapter introdução à mineração de dados com aplicações em ciências ambientais e espaciais. *Instituto Nacional de Pesquisas Espaciais*, p. 15–38, 2008. Citado na página 20.

SARGIANI, V. et al. Identificação de padrões em textos de mídias sociais utilizando redes neurais e visualização de dados. Universidade Presbiteriana Mackenzie, 2018. Citado 3 vezes nas páginas 21, 22 e 31.

SASSI, R. J. *Uma arquitetura híbrida para descoberta de conhecimento em bases de dados: teoria dos rough sets e redes neurais artificiais mapas auto-organizáveis*. Tese (Doutorado) — Universidade de São Paulo, 2006. Citado 2 vezes nas páginas 29 e 30.

SILVA, L.; SILVA, L. Fundamentos de mineração de dados educacionais. In: . [S.l.: s.n.], 2014. p. 568. Citado 2 vezes nas páginas 20 e 21.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. [S.l.]: Elsevier Brasil, 2017. Citado na página 22.

SILVA, M. A. S. da. Mapas auto-organizáveis na análise exploratória de dados geoespaciais multivariados. *SILVA*, v. 681, p. 019, 2004. Citado na página 29.

SMOLENSKY, P. *Connectionism, constituency, and the language of thought*. [S.l.]: University of Colorado at Boulder, 1988. Citado na página 24.

WARE, C. *Information Visualization: Perception for design*. second. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. Citado na página 31.

ZUCHINI, M. H. et al. Aplicações de mapas auto-organizáveis em mineração de dados e recuperação de informação. [sn], 2003. Citado 2 vezes nas páginas 28 e 29.

Appendices