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

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

2.1.1 Classificação

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.

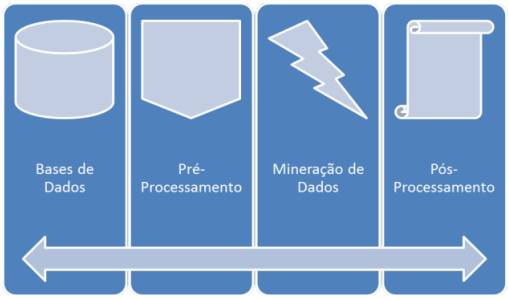
A descoberta de padrões constitui-se de um processo que se inicia pela escolha dos dados que documentam de alguma maneira a pergunta que o especialista deseja responder. Os dados são integrados e pré-processados para que sejam entregues estruturados, higienizados, selecionados e padronizados à tarefa 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.

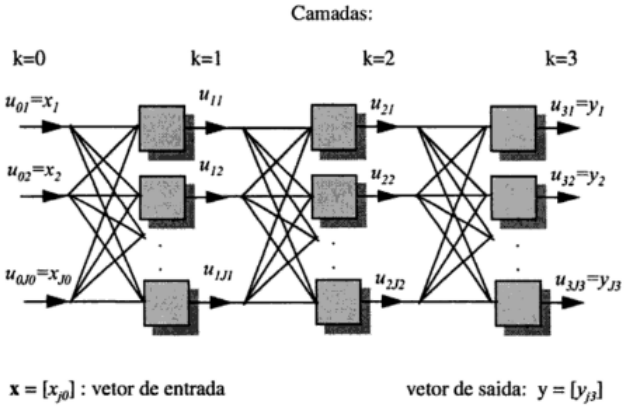
Comforme Braga, Carvalho e Ludermir (2000) e Kovács (2002), por volta do fim da década de 1950, na Universidade de Cornell, Rosenblatt deu continuidade às idéias de McCulloch. Criando uma genuína rede de múltiplos neurônios do tipo *discriminadores lineares* esta rede foi descrita como rede de *perceptron*. Um perceptron é uma rede com a seguinte topologia, os neurônios são dispostoos em váiras *camadas*. Os que recebem das entradas diretamente formam o que é chamada de *camada de entrada*. A camada que recebe a saída da camada de entrada como entrada constituem a segunda camada e assim consecutivamente até a ultima camada que é chamada de *camada de saída*. Camadas que ficam entre as de entrada e saída são comumente referidas como *camadas ocultas*.

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)

Ainda segundo Kovács (2002) e Braga, Carvalho e Ludermir (2000) 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).

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

Figura 2 – Rede Neural Multicamada



Fonte: Kovács (2002, p.40)

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 Mapas Auto Organizáveis de Kohonen

3 Desenvolvimento

4 Testes e Análise de Resultados

5 Conclusão

Referências

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 19.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, v. 5, n. 2, 2006. Citado na página 21.

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 20.

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 2 vezes nas páginas 24 e 25.

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 19.

CECILIO, S.; CASTRO, R. Análise de dados não estruturados: Mineração de textos. In: _____. [S.l.: s.n.], 2015. p. 79–98. Citado na página 20.

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 19.

FERNANDES, A. M. da R. *Inteligência Artificial: noções gerais*. [S.l.]: Visual Books, 2005. Citado na página 22.

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

HODGES, A. *Turing um filósofo da natureza*. [S.l.]: Unesp, 1999. Citado 2 vezes nas páginas 22 e 23.

KOVÁCS, Z. L. *Redes neurais artificiais*. [S.l.]: Editora Livraria da Fisica, 2002. Citado 2 vezes nas páginas 24 e 25.

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 20.

RUSSELL, S.; NORVIG, P. *Inteligência Artificial. Tradução da Terceira Edição*. [S.l.]: Editora Elsevier, 2013. Citado 2 vezes nas páginas 22 e 23.

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 19.

SARGIANI, V. et al. Identificação de padrões em textos de mídias sociais utilizando redes neurais e visualização de dados. Universidade Presbiteriana Mackenzie, 2018. Citado 2 vezes nas páginas 20 e 21.

SILVA, L.; SILVA, L. Fundamentos de mineração de dados educacionais. In: . [S.l.: s.n.], 2014. p. 568. Citado 2 vezes nas páginas 19 e 20.

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 21.

SMOLENSKY, P. *Connectionism, constituency, and the language of thought*. [S.l.]: University of Colorado at Boulder, 1988. Citado na página 23.

Appendices