UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO" – UNESP

FACULDADE DE ARQUITETURA, ARTES E COMUNICAÇÃO - FAAC

PROGRAMA DE PÓS-GRADUAÇÃO DO MESTRADO PROFISSIONAL EM MÍDIA E TECNOLOGIA – PPGMiT

RICHERLAND PINTO MEDEIROS

INFERÊNCIA DE EMOÇÕES EM FRAGMENTOS DE TEXTOS OBTIDOS DO FACEBOOK

Bauru – SP
2017

RICHERLAND PINTO MEDEIROS

INFERÊNCIA DE EMOÇÕES EM FRAGMENTOS DE TEXTOS OBTIDOS DO FACEBOOK

Trabalho de Conclusão de Mestrado apresentado ao Programa de Pós-graduação em Mídia e Tecnologia, da Faculdade de Artes, Arquitetura e Comunicação - FAAC, Universidade Júlio de Mesquita Filho - UNESP, para obtenção do título de Mestre em Mídia e Tecnologia sob a orientação do Professor Titular Dr. João Fernando Marar.

Bauru – SP
2017

Medeiros, Richerland Pinto.
Inferência de Emoções em Fragmentos de Textos Obtidos do Facebook / Richerland Pinto Medeiros, 2017

60 f.

Orientador: João Fernando Marar

Dissertação (Mestrado)–Universidade Estadual Paulista. Faculdade de Arquitetura, Artes e Comunicação, Bauru, 2017

1. Emoções; 2. Processamento de Linguagem Natural; Aprendizado de Máquina, 3. Maximização de Entropia. I. Universidade Estadual Paulista. Faculdade de Arquitetura, Artes e Comunicação. II. Título.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE RICHERLAND PINTO MEDEIROS, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MÍDIA E TECNOLOGIA, DA FACULDADE DE ARQUITETURA, ARTES E COMUNICAÇÃO - CÂMPUS DE BAURU.

Aos 27 dias do mês de abril do ano de 2017, às 10:30 horas, no(a) Sala de Reuniões da Seção Técnica de Pós-graduação da Faculdade de Arquitetura, Artes e Comunicação, reuniu-se a Comissão Examinadora da Defesa Pública, composta pelos seguintes membros: Prof. Titular JOAO FERNANDO MARAR - Orientador(a) do(a) Departamento de Computação da Faculdade de Ciências do câmpus de Bauru / UNESP/ Câmpus de Bauru, Prof. Adj. Dr. ANTONIO CARLOS SEMENTILLE do(a) Departamento de Computação / UNESP-Câmpus de Bauru, Professor Dr. RODRIGO HOLDSCHIP do(a) Centro de Ciências Exatas / Universidade do Sagrado Coração, Centro de Ciências Exatas e Sociais Aplicadas., sob a presidência do primeiro, a fim de proceder a arguição pública da DISSERTAÇÃO DE MESTRADO de RICHERLAND PINTO MEDEIROS, intitulada **Inferência de Predominância de Emoções em Textos**. Após a exposição, o discente foi arguido oralmente pelos membros da Comissão Examinadora, tendo recebido o conceito final: Aprovado. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelos membros da Comissão Examinadora.

Prof. Titular JOAO FERNANDO MARAR

Prof. Adj. Dr. ANTONIO CARLOS SEMENTILLE

Professor Dr. RODRIGO HOLDSCHIP

DEDICATÓRIA

Dedico este trabalho as minhas amadas esposa e filha, por toda paciência, apoio e motivação e aos meus pais que me deram os direcionamentos certos que me ajudaram a chegar até aqui.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por seu imenso amor e sabedoria e por todas as oportunidades que tive até aqui.

Agradeço a minha amada esposa Mônica, por toda sua paciência, equilíbrio e amor não só pelo percurso trilhado no mestrado, mas por todas as fases que me trouxeram até aqui. Com certeza devo muito a ela por todos os momentos de carinho e conversas que me mantiveram firme em todos os momentos de dificuldade.

Agradeço profundamente minha amada filha Helena, pois mesmo sem ao menos entender sua importância, me deu toda motivação para nunca desistir de meus ideais e meus sonhos durante esta caminhada.

Agradeço aos meus pais, por todo amor e dedicação que sempre tiveram e também por todas as correções que recebi ao longo de minha vida.

Agradeço minha sogra Denir, que me acolheu como a um filho e me tratou com muito amor, obrigado por todo o empenho em nos ajudar.

Agradeço ao meu brilhante orientador Fernando Marar por todos os ensinamentos e direcionamentos ao longo desta pesquisa. Devo muito do meu interesse por inteligência artificial a essa grande pessoa.

Agradeço ao professor Denis Renó, por me surpreender em suas aulas, principalmente me mostrando como uma disciplina densa pode ser ministrada de forma leve.

Agradeço a Professora Maria Cristina Gobbi, pela carinho e compreensão durante todos os momentos que estivemos juntos, por toda sabedoria e principalmente pela habilidade que tem de escutar verdadeiramente. Professora, com certeza você deixou uma marca muito positiva em minha vida.

Agradeço a Glaucia Lopes, que sempre me “puxou a orelha” para me formar e que sempre me disse o que eu precisava ouvir.

Agradeço a meus amigos Conrado, Marcos Flávio e Glauco por todos os momentos de risada e por todo incentivo, com certeza eu não seria a mesma pessoa sem vocês.

Agradeço a meus amigos Luiz Gustavo e Rodrigo Fernandes, por sempre me ouvirem e sempre me ajudarem em tudo que precisei.

Enfim, tive a sorte de ter grandes pessoas ao meu lado.

MEDEIROS, Richerland Pinto. **Inferência de Predominância de Emoções em Texto obtidos pelo Facebook**. 2017. Trabalho de Conclusão de Mestrado (Mestrado Profissional em Mídia e Tecnologia) – FAAC – UNESP, sob a orientação do Professor Titular Doutor João Fernando Marar, Bauru, 2017.

RESUMO

Esta pesquisa tem como objetivo analisar o uso da técnica estatística de aprendizado de máquina Maximização de Entropia, voltado para tarefas de processamento de linguagem natural na inferência de emoções em textos obtidos da rede social *Facebook*. Foram estudados os conceitos primordiais das tarefas de processamento de linguagem natural, os conceitos inerentes a teoria da informação, bem como o aprofundamento no conceito de um modelo entrópico como classificador de textos. Os dados utilizados na presente pesquisa foram obtidos de textos curtos, ou seja, textos com no máximo 500 caracteres. A técnica em questão foi abordada dentro do aprendizado supervisionado de máquina, logo, parte dos dados coletados foram usados como exemplos marcados dentro de um conjunto de classes predefinidas, a fim de induzir o mecanismo de aprendizado a selecionar a classe de emoção mais provável dado o exemplo analisado. O método proposto obteve índice de assertividade médio de 90%, baseado no modelo de validação cruzada.

Palavras-chave: Emoções; Processamento de Linguagem Natural; Aprendizado de Máquina, Maximização de Entropia.

ABSTRACT

This research aims to analyze the use of entropy maximization machine learning statistical technique, focused on natural language processing tasks in the inferencing of emotions in short texts from *Facebook* social network. Were studied the primary concepts of natural language processing tasks, IT intrinsic concepts, as well as deepening the concept of Entropy model as a text classifier. All data used for this research came from short texts found in social networks and had 500 characters or less. The model was used within supervised machine learning, therefore, part of the collected data was used as examples marked within a set of predefined classes in order to induce the learning mechanism to select the most probable emotion class given the analyzed sample. The method has obtained the mean accuracy rate of 90%, based on the cross-validation model.

Keywords: Emotions; Natural Language Processing; Machine Learning; Entropy Maximization.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

1 INTRODUÇÃO

Emoções, enquanto objeto de pesquisa, podem ser abordadas por diferentes áreas como a psicologia, a psiquiatria, a neurologia e outras ciências voltadas ao estudo do comportamento e das interações humanas. A emoção consiste de um mecanismo elementar extremamente importante e caracterizador da natureza e da conduta humana, e que, recentemente, tem atraído a atenção de pesquisadores do campo da inteligência artificial, principalmente aqueles ligados aos estudos específicos do processamento das linguagens naturais.

Por isso, a presente pesquisa se desenvolve no âmbito da análise e da identificação de emoções, no âmbito do processamento de linguagem natural, e mais especificamente na subárea da classificação de informações textuais.

Sua relevância fica evidente na medida em que é perceptível o crescimento do fluxo de dados criado diariamente na internet, subsidiado pelo advento das redes sociais, que tem feito emergir um grande volume de textos produzidos diariamente, em *blogs*, *microblogs*, comentários, fóruns e etc. Tal conteúdo configura-se geralmente como meio de expressão de opinião, especialmente nos *microblogs*, fontes ricas em opiniões e emoções na internet (PAK e PAROUBEK, 2010).

Neste sentido, pesquisas voltadas ao processamento de linguagem natural e à descoberta de padrões textuais têm se desenvolvido rapidamente, e com seus avanços, questões subjetivas como análise de emoções têm ganhado grande importância neste campo do conhecimento.

Desta forma, a presente pesquisa apresenta uma abordagem para a inferência de emoções em textos, buscando elucidar conceitos necessários para a construção de um mecanismo de identificação e análise de emoções de textos curtos[1].

[1] Textos curtos serão considerados nesta pesquisa, como textos de no máximo 500 caracteres provenientes de redes sociais.

1.1. Objeto de pesquisa

O objeto da presente pesquisa é o processamento de linguagem natural utilizado na tarefa de classificação de textos, e direcionado ao processo de inferência de predominância de emoções em textos provenientes de redes sociais:

> Basicamente se entende por inferência aquilo que se usa para estabelecer uma relação não explicita no texto, entre dois elementos desse texto. As inferências surgem de uma necessidade e do conhecimento de mundo do leitor (ouvinte). (KOCH e TRAVAGLIA, 1997, p. 70).

Ao inferir emoções em textos é importante notar que as emoções geralmente não ocorrem isoladamente, ou seja, um fragmento textual pode conter alto teor da emoção raiva, somado a um teor menor da emoção medo, por exemplo. Picard (1997) evoca o mecanismo generativo como o fator chave de coexistência das emoções.

1.2. Objetivos gerais e específicos

O objetivo principal deste trabalho é desenvolver um mecanismo de classificação automática utilizando do aprendizado de máquina, mais especificamente no campo do processamento da linguagem natural, por meio da técnica de raciocínio estatístico de maximização de entropia, a fim de inferir emoções a partir de textos obtidos em redes sociais.

Ainda, os objetivos específicos deste trabalho são:

a) Classificar emoções em textos curtos, considerando as 6 emoções primárias: raiva, nojo, medo, alegria, tristeza e surpresa;
b) Analisar a eficiência de técnicas estatísticas de aprendizado de máquina na tarefa de classificação.

1.3. Justificativa

Ano após ano, o crescimento do volume de dados trafegados aumenta em saltos significativos. Segundo a empresa Cisco (2015) a previsão de trafego de dados em 2014 foi de aproximadamente 718,8 *hexabytes*, enquanto a previsão para 2015 foi de 868,8 *hexabytes*, um aumento de 20,87% em relação ao ano anterior. Já a previsão para 2016 era de que o tráfego de dados pela primeira vez ultrapassasse a marca de 1 *zettabytes*, medida que equivale a um

sextilhão de bytes, ao chegar em 1,06 *zettabyte*. Isto significa que, durante o ano de 2016 trafegar-se-ia uma quantidade de dados tamanha que poderia haver, como resultado disto, uma soma equivalente ao que foi trafegado de 1984 a 2012 (CISCO, 2015).

Estes números demonstram o crescimento da produção de dados na internet, uma vez que com seu advento, bem como das redes sociais que nela surgiram, se proliferaram e se expandiram informações e dados ao redor do mundo, a humanidade ganhou uma tecnologia decisiva na era da informação (CASTELLS, 2014) e com isso um dos mais expressivos meios de comunicação, somado à larga oferta de aparelhos digitais provenientes do crescente desenvolvimento tecnológico, tornou viável a conexão à internet para grande parte da população, quer por meio de computadores, *smartphones* ou *tablets*, entre outros.

A popularização da rede global serviu como um convite para a expressão, onde os usuários, além de possuir acesso a informações que consomem, também produzem com frequência conteúdos informativos diversos, em formatos como textos, áudios e vídeos.

Este espantoso volume de informação é produzido diariamente por indivíduos com variados perfis e motivações. Os dados gerados por estes usuários estão separados, para os fins desta pesquisa, em dois grupos: os estruturados e os não estruturados Chamamos dados estruturados aqueles que recebem algum tipo de estrutura e a obtenção de informação a partir deles é fácil e objetiva, enquanto os dados desestruturados são aqueles que não possuem estrutura óbvia, ou seja não há separação de informações por tipos, ou predefinições, a exemplo textos puros escritos em linguagem natural, vídeos, músicas e etc. O volume de dados que mais cresce é exatamente o de dados desestruturados e neles estão os maiores desafios de processamento de linguagem natural e identificação de emoções.

Isto por que o surgimento das redes sociais, trouxe consigo um interessante fenômeno: todos em qualquer momento podem opinar e ter sua opinião amplamente comentada e divulgada, ou seja, seu advento deu voz pública a quem não a tinha e trouxe à luz da esfera pública ilustres desconhecidos. Assim, diariamente é produzido um grande volume de opiniões em texto ou audiovisual nas redes sociais sobre os mais diversos acontecimentos, pessoas, produtos e marcas, configurando-se um grande volume de informação gerado a respeito de gostos e preferências dos mais variados públicos, o que acaba por chamar a atenção de empresas e marcas que enxergam grandes oportunidades de estabelecer e ampliar sua competitividade e de entender melhor as necessidades de seus consumidores.

Neste momento, é importante perceber que surgem indagações pertinentes a este cenário de dados, de forma a viabilizar sua utilização e seu entendimento:

- Como transformar os dados gerados por todas essas interações a partir do texto puro em informações, quantificáveis e tabuláveis?
- Além disso, como entender a regionalidade das opiniões e preferencias?

Com isso, a análise de linguagem natural sempre se mostrou uma tarefa complexa demais para soluções em tempo algorítmico. Assim, o uso de técnicas de aprendizado de máquina, mais especificamente tarefas de processamento de linguagem natural tem sido usadas para tentar sanar estas questões.

Portanto, temos que, com a evolução e a ampliação significativa do uso das redes sociais, faz-se premente a evolução no uso de tecnologias que possam amparar o entendimento dos dados gerados diariamente nestes ambientes tecnológicos dispostos.

Ou seja, se tomamos a necessidade de processamento, entendimento e análise de conteúdos e dados gerados em rede, temos que estes fenômeno se caracteriza como um dos grandes impulsionadores das pesquisas acerca do processamento de linguagem natural e de sua galopante evolução nos dias de hoje. Neste contexto se justifica a realização de uma pesquisa voltada a inferir emoções provenientes de redes sociais, com base nas emoções primárias, segundo Damásio (2000), pois a identificação e a análise das emoções vai além da obtenção de informações superficiais, evidenciando o cerne dos motivadores da produção textual.

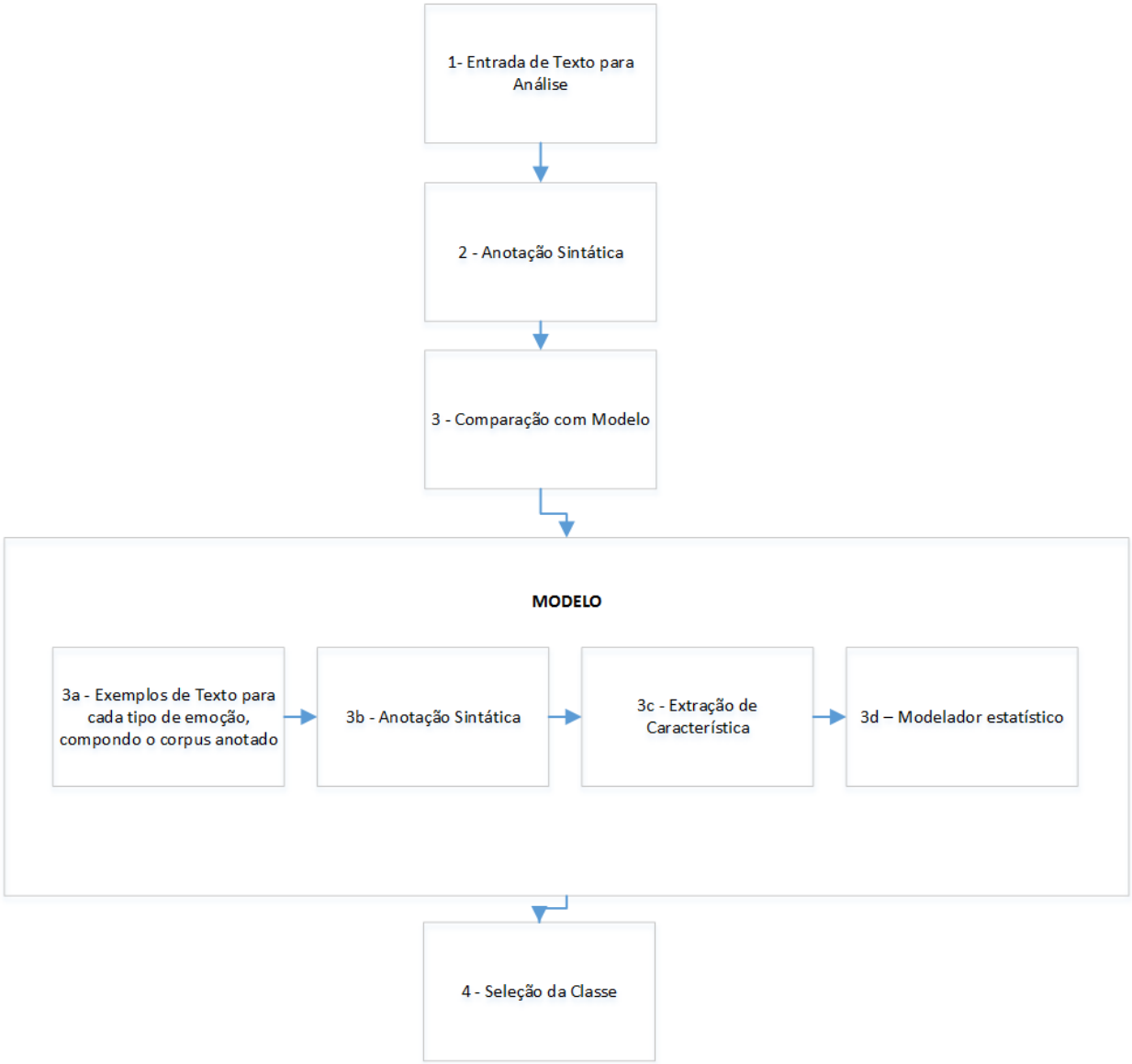
1.4. Plano metodológico

Para este trabalho foi realizada uma revisão bibliográfica acerca do método de classificação de texto por maximização de entropia (JAYNES, 1957), constituída a partir do aprendizado supervisionado de máquina, usando como treinamento fragmentos textuais previamente classificados por interação humana.

Desta forma, foram consideradas como classes as emoções primarias defendidas por Damásio (2000) e Paul Ekman (EKMAN e FRIESEN, 1978). A priori foi desenvolveu-se um protótipo para testar a hipótese da inferência de emoções. Os materiais (Figura 1) envolvidos na identificação de emoções por meio da classificação, foram testados a fim de quantificar índices de acerto baseados nas métricas: Precisão, Cobertura e Medida-F, considerando a comparação com um conjunto de dados de controle.

O protótipo desenvolvido seguiu o seguinte fluxo:

Figura 1 - Fluxo Macro de Execução



Fonte: Produção Própria.

1 – Entrada de Texto para Análise: Submissão do texto a ser classificado, obtido por meio de redes sociais;

2 – Anotação Sintática: Uso de informações sintáticas do texto em questão para a obtenção de características;

3 – Comparação com Modelo: Uso da técnica de maximização de entropia amparado por um corpus anotada, induzindo as características de cada emoção para a seleção da classe candidata;

3a - Exemplos de Texto para cada tipo de emoção, compondo o corpus anotado: Cada classe (ou tipo de emoção) possui uma quantidade de 15 fragmentos de textos curtos anotados previamente, para a seleção de características principais;

3b – Análise Sintática: O modelador de maximização de entropia usa também na leitura do corpus anotado a análise sintática para a obtenção de características;

3c – Extração de Características: todas as características são tabuladas no formato de distribuição de probabilidades e servirá como substrato principal para o modelador de maximização de entropia;

3d – Modelador Estatístico: É aplicado a maximização de entropia para selecionar a probabilidade das classes serem escolhidas como candidatas dado o modelo proposto;

4 – Seleção da Classe: Após o retorno do modelador estatístico é selecionada a classe com maior probabilidade considerando o texto analisado.

As macro etapas da pesquisa consistiram da revisão da literatura científica sobre classificação de textos; desenvolvimento de um sistema (protótipo) voltado à classificação de emoções; análise dos resultados obtidos por meio da utilização do protótipo; apresentação de uma matriz de resultados considerando medida-f, precisão e cobertura.

1.5. Organização da dissertação

A estrutura do presente trabalho organiza-se, no primeiro capítulo após a introdução (2) em torno do levantamento teórico acerca de processamento de linguagem natural e seus conceitos primordiais com foco na tarefa de classificação de textos, resultando na inferência de emoções, bem como a aplicação destas técnicas em um protótipo de classificação com maximização de entropia e o estudo sobre as emoções, linguagem e comunicação inseridos no contexto de aprendizado de máquina.

O capítulo seguinte (3), traz uma coletânea de conceitos acerca de processamento de linguagem natural e aprendizado de máquina, visando contextualizar as ferramentas utilizadas no processo de identificação de emoções, possibilitando a inferência de suas predominâncias. Assim, este capítulo apresenta uma discreta revisão acerca da emoção e sua definição, separando-a do conceito do sentimento, explorando os princípios e características deste processo neurobiológico, com vistas a ampliar possibilidades de estudo considerando o campo de processamento de linguagem natural.

O capítulo posterior (4) segue apresentando conceitos importantes para o presente trabalho no que se refere aos métodos de aprendizado de máquina e suas bases teóricas, como o conceito da entropia na teoria da informação, voltado a quantificação de informação e a técnica maximização de entropia, fundamentada na teoria da informação.

O capitulo (5) apresenta a implementação do mecanismo de classificação de textos utilizado no processo de inferência de emoções, dissertando sobre as etapas necessárias para a construção, utilização e analise de resultados do modelo de generalização proposto. O capitulo (6) apresenta os resultados da pesquisa, considerando a metodologia implementada e os objetivos da presente pesquisa. Por fim as Conclusões (7) são apresentadas, bem como proposta de trabalhos futuros.

2 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA

O presente capítulo apresenta um conjunto de conceitos acerca do aprendizado de máquina, e sendo que o objeto geral da presente pesquisa o processamento de linguagem natural, enquanto subárea deste campo de pesquisa, neste trabalho propõe-se um método de inferência de emoções utilizando o aprendizado estatístico de máquina, ao especificar o processo de aprendizado estatístico determina-se que o modelo de raciocínio de máquina ocorre devido a processos de indução estatística. Existem outros paradigmas de raciocínio para o aprendizado de máquina, porém não serão foco do presente trabalho.

2.1. Inteligência artificial

Entendemos que o conceito de inteligência artificial remete ao conjunto de técnicas e mecanismos tecnológicos, usados com objetivo de "imitar" inteligência cujo conceito compreendemos a partir das características humanas de aprendizado.

Esta área de estudos e pesquisa da Ciência da Computação pode, de maneira bastante superficial, ser dividida em duas ramificações: a da Inteligência Artificial Forte e a da Inteligência Artificial Fraca. O termo Inteligência Artificial Forte foi cunhado originalmente pelo pesquisador John Irvin Good (1965) e apesar de trazer questões éticas à tona quando de sua pesquisa, sua real possibilidade não se configura como unanimidade entre os pesquisadores da área, pois trata-se da inteligência artificial consciente, ou seja, mecanismos tecnológicos que, de fato, podem pensar e sentir como um indivíduo humano, de modo que o processo seja indistinguível daquele de um ser humano. Já o termo Inteligência Artificial Fraca busca resolver problemas complexos demais para que sejam tratados em tempo algoritmo, e este campo de estudo não se preocupa se um agente inteligente possui entendimento ou questão de problema, mas sim, se dado um conjunto de instruções e um acumulado de conhecimento possa ser generalizado e inferido a fim de obter resultados e ações para a solução de problemas.

2.2. Aprendizado de máquina

Já o termo "aprendizado de máquina" refere-se a uma subárea do campo da Inteligência Artificial que estuda formas e meios de proporcionar a mecanismos

computacionais a capacidade de aprender, com base no conhecimento prévio ou na experiência adquirida por meio de supervisão ou desenvolvimento do processamento de um problema. O grande objetivo, neste caso, é resolver problemas que sejam complexos demais para soluções em tempo algorítmico (WEISS e KULIKOWSKI, 1991).

Os estudos acerca desse assunto são separados, essencialmente, em aprendizado supervisionado e aprendizado não-supervisionado. O primeiro se refere ao aprendizado baseado em treinamento prévio de um modelo de generalização para a solução de um problema, ou seja, requer que um indivíduo determine como deve ser o funcionamento do sistema para a obtenção de um resultado; este processo é voltado ao processamento de linguagem natural, e a supervisão é geralmente ligada à anotação de fragmentos de informações dentro de bases textuais, que compõem o modelo de generalização, ou na atribuição de exemplos de fragmentos textuais a classes predeterminadas. Já o segundo não requer supervisão ou treinamento, usando técnicas que separam as informações com base em características intrínsecas ao problema observado em processamento de linguagem natural. Assim, geralmente é um processo utilizado para a criação de *clusters* em que se busca separar linearmente problemas, distinguindo-os em dois ou mais clusters, com base na dicotomia de pertinência ou não pertinência. O raciocínio não supervisionado descobre as relações e fenômenos estatísticos presentes, por meio de características observadas no próprio problema.

Considere-se, ainda, que existem alguns paradigmas de funcionamento inseridos no aprendizado de máquina, os quais descrevem o modelo ou simplesmente os pressupostos que um algoritmo ou conjunto de algoritmos inteligentes deve adotar para compor uma solução inteligente. Os principais paradigmas, a saber, simbólico, conexionista e estatístico, podem ser definidos das seguintes formas:

- Paradigma Simbólico é um sistema de representação de informação e se baseia na construção de representações de conceitos por meio da análise de exemplos e suas negações, produzindo simbologia, e são geralmente descritos na forma de expressões lógicas como árvores e regras de decisões, redes de informações e etc.
- Paradigma Estatístico é o que se ampara em processos e técnicas estatísticas, envolvidos na obtenção e detecção de características e na compreensão do funcionamento da informação disposta no problema, e estuda os fenômenos estatísticos inerentes ao contexto do problema;
- Paradigma Conexionista ou Conexionismo é um modelo de raciocínio baseado em métodos matemáticos em funcionamento em uma estrutura espirada em redes neurais humana.

> Esses modelos (conexionistas) assumem que o processamento de informação ocorre pela interação de um grande número de elementos processadores simples chamados de unidades, cada um enviando sinais excitatórios e inibitórios para os outros (RUMELHART et al., 1986, p. 98).

Existem outros paradigmas de aprendizado de máquina, como o genético, baseado em instâncias, entre outros que não serão investigados durante o presente trabalho, devido a escolha do paradigma estatístico no uso da técnica de maximização de entropia.

2.3. Processamento de linguagem natural

Entendemos que o processamento de linguagem natural consiste de uma importante tarefa do campo de inteligência artificial, intimamente ligada ao estudo de linguística computacional e a seu objeto no âmbito da problemática de interpretação, extração e criação textual automática de linguagem humana.

Os estudos desenvolvidos neste campo viabilizam a propícia interpretação, análise ou extração de grandes volumes de dados, especialmente quando o volume dos dados se configura como uma barreira para a análise humana direta. A interpretação está intimamente relacionada com a semântica, ou seja, o estudo do sentido ou significado das palavras (POTTIER, 1978). Ainda que não haja clareza em como representar significado ou sentido de palavras dentro do contexto computacional, Chaudiron (2007) discorre sobre a impossibilidade de atribuição de significado ao conteúdo, porem reforça a possibilidade da análise das relações válidas entre as palavras, a partir de seus conceitos.

Sendo o estudo do significado um assunto amplamente debatido, percebemos a partir deste que existem formas de representação que visam a aproximação do conceito de significado como as ontologias e outros recursos, tratando-se, todavia, de um tema longe de ser completamente resolvido.

A tarefa de processamento de linguagem natural pode utilizar dispositivos simples como N-gramas, ou seja, separação de frases e fragmentos de N Palavras (JURAFSKY e MARTIN, 2008), quanto técnicas mais elaboradas como abordagem matemáticas e estatísticas como Modelos Ocultos de Markov (NORVIG e RUSSEL, 2003).

A seguir, apresentamos as fases específicas dentro do presente contexto.

2.3.1. Pré-processamento

Fase de tratamento do texto “cru”, ou seja, sem qualquer alteração, bem como indexação das palavras quando necessário e normalização das informações contidas nelas.

Nessa fase é realizada a filtragem de *stopwords* (lista de palavras com baixa relevância contextual, as quais serão removidas do texto por não demostrarem expressividade em análises).

Segundo Palmer (2010) a etapa do pré-processamento é de extrema importância para uma análise eficiente, considerando o tratamento de qualquer problema de análise de linguagem natural.

2.3.2. Análise léxica

A análise léxica ou morfológica trata a forma das palavras que a serem analisadas, ou seja, está embasada na construção das palavras, a que Hippisley (2010) compara a importância na construção de texto aos tijolos em uma edificação; tal análise visa prever um conjunto de prefixos, sufixos e radicais e a forma de construir as palavras.

Esta etapa tem grande relevância, pois evita a repetição desnecessária de palavras na construção de dicionários e otimiza o processo de uso dos mesmos, além de auxiliar no processo de obtenção de relevância de termos, uma vez que a análise dos fragmentos pode ser feita com base nos radicais, tomando em conta a maior incidência da presença de um termo radical.

2.3.3. Análise sintática

Análise sintática é a que trabalha com a estrutura das sentenças e orações em um sistema textual. Esta é uma tarefa essencial no processo uma vez que o significado não é encontrado isoladamente em uma palavra mais sim em uma frase (LJUNGLOF e WIRÉN, 2010).

Nesta fase são estudados os sintagmas nominais e os sintagmas verbais. Toda frase é formada por unidades de significado e a estas é dado o nome de sintagmas. Sintagma nominal é referido quando o núcleo do significado é articulado sob o substantivo, enquanto o sintagma verbal é articulado no verbo.

A análise sintática é, talvez, a mais representativa dentro dos estudos e pesquisas de processamento de linguagem natural (DALE, 2010), uma vez que nesta etapa seja realizada a anotação gramatical de cada palavra ou expressão, também se perfaz o enquadrando dentro de uma das 10 classes gramaticais da língua portuguesa.

2.3.4. Análise semântica

De modo geral, podemos afirmar que palavras sozinhas não fazem sentido para o processamento dos computadores, pois eles, nativamente, só compreendem código binário. A análise semântica visa, por meio de um conjunto de técnicas, inclusive fundamentado nas etapas anteriores, inferir o significado de sentenças e palavras dentro de um contexto observado.

Chaudiron (2007) em seu trabalho afirma que não é possível trazer significado diretamente a palavras, mas é possível analisar as relações relevantes e válidas entre as palavras e inferir o significado.

2.3.5. Análise pragmática

Com a obtenção de significado na etapa semântica, a etapa pragmática se encarrega de investigar o funcionamento das sentenças e palavras de acordo com o contexto e o discurso aos quais estão inseridas. Cherpas (1992) a defini como a etapa que dá funcionalidade às palavras dentro de uma linguagem. A análise pragmática constitui uma ferramenta especialmente relevante (MOENS, UYTTENDAELE e DUMORTIER, 1999) uma vez que é fonte de obtenção de características para a desambiguação dentro dos sistemas de processamento de linguagem natural.

2.3.6. Corpus

Um dos recursos mais relevantes dentro do processamento de linguagem natural é o uso de *corpus*, que segundo definição de Sardinha (2004) é uma coleção de dados linguísticos, como textos ou fragmentos de texto, bem como qualquer produto textual dispostos em uma determinada língua, escolhidos baseado em uma necessidade, caracterizando-se como uma amostra linguística.

Os *corpora* se dividem em textos puros ou anotados, em que os primeiros não possuem nenhuma informação de referência, ou seja, nenhuma marcação de classificação ou

identificação; já os textos anotados, por sua vez, possuem um conjunto de documentos marcados, ou seja, com o conhecimento que se quer induzir previamente selecionado dentro do domínio escolhido.

2.3.7. Classificadores

Chamamos de classificadores uma classe de ferramentas do campo de processamento de linguagem natural que utilizam um conjunto de características ou *features* para inferir a pertinência de um determinado fragmento textual a uma classe. Os classificadores são a melhor solução para tarefas de análise de sentimento (CARDOSO, ESTEVES e FONSECA, 2014). Existe um vasto repertório de anotadores, dos quais o modelo de maximização de entropia, ou princípio de máxima entropia, foi selecionado para a presente pesquisa.

A maximização de entropia é uma composição da técnica da estatística indutiva de regressão logística multinominal, somada ao uso de características linguísticas, que tem como objetivo detectar funções e comportamentos produzindo um modelo que permita a indução de valores tomados por uma variável categórica.

Neste caso, entropia é definida como uma medida para incerteza, ou seja, é a quantificação de informação contida em uma variável aleatória. Basicamente a técnica provê a construção de um modelo de distribuição de probabilidade p que se aproxime de p', sendo que p' seja uma distribuição de probabilidade obtida por meio de dados de treinamento (CARVALHO, 2012).

2.3.8. Processamento de linguagem natural e emoções

Existem algumas formas de identificar a presença de emoções em textos, mas duas segundo Plutchik (1980) são as mais comuns:

- As que trabalham com a polarização inferindo como resposta a dicotomia positivo e negativo, geralmente com base em um léxico de palavras e expressões, pontuadas, ao qual cada fragmento (palavras ou expressão) recebe um gradiente numérico definindo o "peso" da emoção; auxiliado por uma técnica estatística para compor a quantificação final da relação de pesos positivos e negativos;
- A segunda, com base na classificação das emoções considerando um rol de possibilidades, ou classes.

Para a presente pesquisa, as classes de emoções adotadas são das emoções primárias, definidas segundo a pesquisa de Damásio (2000), no âmbito da neurologia: alegria, tristeza, medo, nojo, raiva e surpresa, em que exemplos de textos relativos a cada emoção são anotados previamente e uma técnica de descoberta de padrões é utilizada para identificar a qual classe um determinado texto pertence, com base nas características dos textos fornecidos na anotação.

3 EMOÇÕES, LINGUAGEM E COMUNICAÇÃO

Segundo Damásio (2000), emoção é um processo neurobiológico, elementarmente uma variação psíquica e físico-química, iniciada por um ou mais estímulos sensoriais e suas características são experimentadas individualmente como resposta direta ao estimulo recebido. Notadamente emoções são os meios naturais primários de avaliação e coexistência com e no ambiente.

Pinto (2001) complementa afirmando que as emoções envolvem completamente o indivíduo sendo uma reação complexa do corpo e da mente. Goleman (1997) afirma, ainda, a existência de centenas de emoções, considerando combinações, variações, mutações e tonalidades das experiências emocionais, classificando-as como um leque de propensões para a tomada de ações.

De fato, há muitos estudos acerca das emoções, pois sua ubiquidade nos seres humanos encontra vazão nas mais variadas áreas de interesse. Emoções, enquanto objeto de pesquisa, podem ser abordadas por diferentes áreas, como a psicologia, psiquiatria, neurologia e outras ciências voltadas para estudo do comportamento e interações humanas.

Nesta pesquisa, entendemos que a emoção é um mecanismo elementar extremamente importante e caracterizador da natureza e da conduta humana, e que, recentemente, tem atraído a atenção de pesquisadores de inteligência artificial, principalmente aqueles ligados ao processamento de linguagens naturais.

É necessário apontar, neste sentido, que embora emoções possam ser frequentemente confundidas com sentimentos, há uma expressiva diferença entre ambos, apesar de comumente referidos como sinônimos. Emoções são inerentes ao subconsciente, caracterizando-se como uma resposta instintiva ao meio e desencadeada automaticamente como resposta a estímulos. Ou seja, a experiência emocional é interna, particular e parte do que temos de mais primitivo, enquanto os sentimentos se voltam ao raciocínio, à experiência temporal, dado um conjunto de percepções e conclusões, dependendo da compreensão de um conjunto de fatos e acontecimentos, portanto os sentimentos são diretamente ligados as emoções, considerando as respostas emocionais e as conclusões racionais.

Muitos autores em seus estudos subdividem as emoções em categorias, a fim de compreender seus mecanismos e funcionamentos. Damásio (2000) parte de uma perspectiva

neurológica separando as emoções em: primárias, secundárias e emoções de fundo, e as define de forma hierárquica.

Considerando as emoções primárias como inatas e espontâneas, ou seja, independentes de fatores externos, como meio social, cultural ou econômico, consistem de emoções elementares e podem desencadear outras emoções, considerando fatores externos e podemos defini-las como alegria, tristeza, medo, nojo, raiva e surpresa.

Já as emoções secundárias partem de um ajuste do indivíduo aos estímulos socioculturais e variam completamente de acordo com a cultura e os costumes, de modo que inclusive sua percepção externa seja subjetiva e percebida relativamente quando analisada por agentes externos ao meio que as originou.

Por fim, as emoções de fundo possuem um âmbito particular e são inerentes à condição de existência, ou seja, o bem-estar interno, originando-se a partir das considerações das nuances cotidianas, traduzindo-se objetivamente em estado de stress ou relaxamento, disposição ou apatia.

O foco da presente pesquisa calca-se nas emoções primárias: alegria, tristeza, medo, nojo, raiva e surpresa, conforme definidas por Damásio. Entretanto, vale ressaltar que a classificação das emoções ou mesmo suas subcategorias são tema ainda em discussão constante e longe de alcançar um consenso no meio cientifico.

Ekman e Friesen (1978), antes de Damásio, partiram da análise de emoções primárias ou básicas considerando um modelo similar, ambos inspirados nos experimentos de Darwin (1872); posteriormente, Ekman reformulou sua opinião, acrescentando o desdém como emoção primária, aumentando assim para sete emoções básicas.

Abaixo, na figura 2 (ORTONY e TURNER, 1990) encontramos retratada uma coletânea de autores e seus conceitos de emoções primárias.

Figura 2 - A Selection of Lists of "Basic" Emotions - Uma seleção de Listas de Emoções "Básicas".

Reference	Fundamental emotion	Basis for inclusion
Arnold (1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	Relation to action tendencies
Ekman, Friesen, & Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Frijda (personal communication, September 8, 1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James (1884)	Fear, grief, love, rage	Bodily involvement
McDougall (1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mowrer (1960)	Pain, pleasure	Unlearned emotional states
Oatley & Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Panksepp (1982)	Expectancy, fear, rage, panic	Hardwired
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson (1930)	Fear, love, rage	Hardwired
Weiner & Graham (1984)	Happiness, sadness	Attribution independent

Fonte: (ORTONY e TURNER, 1990).

O ponto central está relacionado ao interesse de entender profundamente as emoções primárias e como elas cooperam para a produção de outras emoções, sentimentos ou mesmo colabora no processo cognitivo. Oportunamente, lembramo-nos que Jean Piaget (1896-1980) foi um dos precursores do estudo das emoções e da razão como partes de um mesmo núcleo, afirmando que a afetividade e cognição são diferentes em natureza, porém inseparáveis nas ações humanas.

> É incontestável que o afeto desempenha um papel essencial no funcionamento da inteligência. Sem afeto não haveria interesse, nem necessidade, nem motivação; e consequentemente, perguntas ou problemas nunca seriam colocados e não haveria inteligência. A afetividade é uma condição necessária na constituição da inteligência mas, na minha opinião, não é suficiente.
>
> Podemos considerar de duas maneiras diferentes as relações entre afetividade e inteligência. A verdadeira essência da inteligência é a formação progressiva das estruturas operacionais e pré-operacionais. Na relação entre inteligência e afeto, podemos postular que o afeto faz ou pode causar a formação de estruturas cognitivas. (PIAGET, 1961, p. 52)

A relação entre emoções e cognição é relativamente recente, pois em tempos passados não tão distantes, filósofos e outros pensadores inferiam uma separação clara entre razão e emoção, que pode ser percebida, por exemplo, na celebre frase de René Descartes (1596-1650) “Penso, logo existo”, demostrando o princípio de que a razão e a lógica eram interpretadas hierarquicamente superiores à emoção, fazendo inclusive que a emoção fosse vista como fraqueza ou inferioridade. Anteriormente, Platão (428-347 a.C) já sugerira a abdicação a todos as paixões e prazeres em troca do pensamento, e, no mesmo sentido, Kant (1724-1804) afirmava a impossibilidade da felicidade e a razão coexistirem. Até mesmo os estudiosos estoicos (Sêneca, Epicteto, Marcus Aurelius e etc) travavam disputas em nome da razão, relegando a emoção, o prazer e mesmo o sofrimento como meras experiências externas a serem abandonadas, visto que a razão, o pensamento e a filosofia eram entendidos como superiores a qualquer manifestação afetiva.

> Caso o homem decida identificar o mal que o acomete, iniciando um tratamento, a busca filosófica, em breve descobrirá que não se trata de um exercício difícil: passada a fase inicial, o tratamento deixa de amargar e torna-se mesmo uma fonte de prazer enquanto se processa a cura. Com os remédios do corpo, o prazer só chega depois da cura; a filosofia, pelo contrário, é salutar e saborosa simultaneamente (SÉNECA, 1991, p. 172).

Além de Piaget, Lev Vygotsky e Henri Wallon estudaram a relação entre afeto e cognição (TAILLE, 1993). Para ambos as emoções faziam parte do funcionamento mental e os processos inerentes à mente estavam conectados e, portanto, a razão estava intimamente ligada e dependente das emoções. Wallon, por sua vez, atentou para a predominância intermitente entre razão e emoção dado fatores contextuais. Assim, considerando Piaget, Vygotsky, Wallon, entre outros, a evolução do papel das emoções no teatro da cognição teve grandes acréscimos e abriu frentes de pensamento que iam além do que se tinha como certo pelos grandes filósofos e pensadores.

3.1. Aspectos fisiológicos da emoção

Segundo Maclean (1990) o cérebro humano é divido em três unidades funcionais, a saber: cérebro reptiliano, cérebro dos mamíferos inferiores e cérebro racional. O cérebro reptiliano ou arquipálio é a unidade mais primitiva, sendo responsável pelos sistemas instintivos e reflexos, enquanto o cérebro dos mamíferos inferiores ou paleopálio, além dos recursos de seu precursor (arquipálio) é composto do sistema límbico, responsável pelas emoções e comportamentos emocionais em mamíferos; por último o cérebro racional ou neopálio é

responsável pelos processos mais elaboradores e minuciosos da cognição. O lobo límbico foi nomeado pelo neurologista francês Paul Broca (1878), ao observar uma região cinzenta circular logo abaixo do córtex; esta região é responsável pelas funções emocionais, além dos aspectos de auto identidade e outras funções ligadas a memória, posteriormente Mclean trouxe a denominação sistema límbico.

A emoção é uma função cerebral distribuída, conforme o que demostrou o neuroanatomista James Papez (1937); Papez analisou a interligação nervosa de quatro estruturas básicas a saber: hipotálamo e corpos mamilares (região do paleopálio), núcleo anterior do tálamo (região intermediaria entre o córtex e o paleopálio), giro cingulado (camada de comunicação entre o córtex e o sistema límbico) e o hipocampo (região do córtex, principal responsável pela memória); Maclean mais tarde adotou a proposta de Papez, propondo outras interconexões envolvidas nas emoções.

Os estudos de Papez e MacLean deixaram claro que não há um núcleo isolado responsável pelo processo das emoções e a despeito de algumas áreas terem maior contribuição do que outras no processo a emoções dependem diretamente da interligação das unidades cerebrais. A dicotomia razão-emoção, antes vista como antagônica, ganhou relevância nos estudos da neurociência. A distribuição dos papéis do funcionamento da emoção e a interligação das áreas cerebrais, mostram claramente o papel das emoções na cognição; Damásio (2000) afirma que apesar das emoções não serem mecanismos racionais, desencadeiam processos cognitivos por meio dos sentimentos.

3.2. Comunicação e linguagem

Segundo Chiavenato (2000), comunicação é o processo de transmissão de informação de um indivíduo a outro, processo este inerente à necessidade de convivência em sociedade do ser humano e intimamente ligado ao conceito de sociedade, pois sem a comunicação não existe interação entre os indivíduos e também não há a perpetuação dos padrões e processos culturais.

Já segundo Fischer (1983) expressar-se é uma necessidade humana, e por meio da comunicação os indivíduos podem, além de interagir com o meio, influenciar mudanças e criar novos hábitos e costumes.

A toda comunicação é inerente um agente emissor que carrega consigo uma mensagem, independente do código ou mecanismo de codificação da mensagem; seu objetivo é chegar a um receptor, fechando assim seu ciclo, e a partir destas considerações, Rudger (1998)

postula que a comunicação é transmitir a maior quantidade de informação em menor tempo, com a maior fidelidade a mensagem original possível. Com a evolução da espécie humana o modelo de comunicação aperfeiçoou-se e as linguagens tornaram-se complexas e com regras bem definidas.

> A comunicação envolve o significado ou a interpretação das mensagens, que dependerá da dimensão semântica do código ao qual está referido. As mensagens só adquirem sentido quando são rebatidas a códigos, e a atualização deste dá-se por meio das mensagens. A informação depende apenas da variedade ou do número de mensagens possíveis abrangidas pelo código. (EPSTEIN, 1988, p. 16).

A linguagem é o código conhecido e compreendido tanto pelo emissor quanto pelo receptor, dotando-os de discernimento, fazendo com que por meio dela, a linguagem, o receptor receba a mensagem originada do emissor, de forma que em seu cérebro humano isso implicará no processo de interpretação de pensamentos e raciocínios, (re)formulando a mensagem entregue de forma a ele compreensível.

Assim como as emoções, a linguagem não possui um único núcleo responsável por seu controle e construção, portanto ela é distribuída em áreas diversas que vão do sistema límbico, passando pelo lobo occipital. Contudo, duas áreas são especialmente importantes, quando a linguagem é falada e escrita, a saber a área de Brocca, situada no córtex pré-frontal responsável diretamente pelo núcleo formador das palavras, e a área de Wernickie, situada no lobo temporal, servindo basicamente como um processador de sons e padrões, que possibilitam a interpretação das ondas sonoras como palavras por nosso cérebro (KONKIEWITZ, 2009).

Vale a reflexão de que a comunicação se dá por meio da linguagem e esta é inerente a processos mentais cognitivos e ainda amparada pela visão de Damásio (2000), cujo ensaio acerca das emoções conduz o conceito de cognição.

O processo linguístico é subproduto das emoções, desde a necessidade afetiva da comunicação e interação entre indivíduos até os mecanismos pelos quais a linguagem é produzida e usada no cérebro humano e por consequência se transforma também em um meio de expressão das emoções.

Desta forma, fica perceptível que a linguagem produzida em textos além de informações, contém a expressão de emoções e consequentemente sentimentos e opiniões, e nos cabe ressaltar que entender ou interpretar quais emoções exatamente estão contidas em quais fragmentos de texto é uma tarefa complexa; mesmo as mais elaboradas estratégias de identificação de emoções, que podem produzir ambiguidade (CUCS, 2005), pois cada

indivíduo possui características emocionais próprias, atribuindo emoções pragmaticamente de acordo com as experiências por eles vividas.

4 ENTROPIA DA INFORMAÇÃO E MAXIMIZAÇÃO DE ENTROPIA

Entropia é a medida de desordem de um sistema – esta é, entre outras, a definição para o termo que foi cunhado a mais de 150 anos (MATTOS e VEIGA, 2002), e o conceito se exprime em uma palavra cuja raiz etimológica origina-se no grego antigo, com os radicais *en* – (em, sobre, perto de, etc.) e *tropêe* (mudança, o voltar-se, alternativa, troca, evolução, entre outros.). O termo foi primeiramente usado pelo físico alemão Rudolf Julius Emmanuel Clausius (1822-1888) e originalmente era utilizado no âmbito dos estudos acerca das leis da termodinâmica, sendo mais tarde difundido em outras áreas da ciência para descrever a quantidade de "bagunça", caos ou desordem em um dado sistema observado.

Já o matemático Claude Shannon, um dos precursores da teoria da informação, usou o termo em seus estudos voltados a quantificar o montante de informação contido em uma mensagem, considerando o processo de entrega saindo do emissor e chegando ao seu receptor. A teoria da informação é o campo da matemática que estuda a quantificação da informação, uma vez que o conceito por si só seja amplo demais para ser definido com precisão, este campo estuda diferentes formas e diferentes perspectivas da informação pelos mais diversos campos da ciência, num esforço interdisciplinar. Uma das definições pertinentes a este campo e que nos será útil nesta pesquisa é a de Kobashi e Tálamo (2003):

> A informação, que se apresenta como objeto de estudo da Ciência da Informação, é uma estrutura significante que sintetiza os conteúdos dos documentos, sob formas diversas, segundo políticas e segmentos de usuários. [...] o valor da informação consiste, conforme já afirmado, em gerar conhecimento. (KOBASHI e TÁLAMO, 2003, p. 19)

Esta definição ilustra o valor que a informação possui ao gerar conhecimento. A teoria da informação se volta a quanta informação é percebida, quanto que um conjunto de dados carregada de informação, considerando as restrições do canal ao qual ela é transportada. Apesar da aparente simplicidade que pode aparentar inicialmente, a teoria da informação contribuiu com uma ampla mudança global no formato de como víamos ou entendíamos a informação (AFTAB, et al., 2005); Shannon nos mostrou como a informação poderia ser medida com absoluta precisão e ainda nos deu mecanismos para avaliar qual unidade ou qual fragmento de uma determinada comunicação podia constituir-se como primordial para o correto entendimento da mensagem. Dentro de sua teoria há quatro conceitos principais e a partir deles é possível entender o porquê do potencial transformador da teoria.

4.1. Capacidade do canal e ruídos de codificação no canal

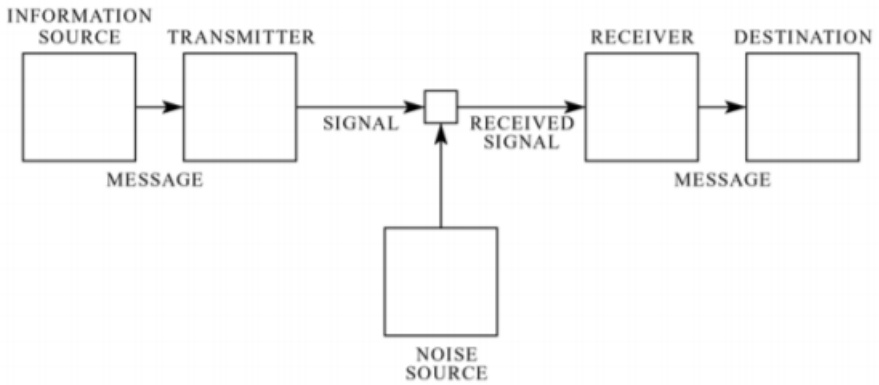
Provavelmente o resultado objetivo mais expressivo tenha sido expresso pelo conceito de que cada canal de comunicação possui limitações e estas devem ser consideradas no processo de transmissão de informação. Originalmente o limite observado foi a velocidade medida em bits por segundo. A investigação é amplamente conhecida como Limite de Shannon.

Por mais simplista que possa parecer este conceito, no âmbito da presente pesquisa, sua aplicação foi essencial, e foi possível realizar a transmissão de informações com zero erros, em decorrência das devidas considerações de limitação do canal para a transmissão de uma mensagem.

4.2. Arquitetura formal de sistemas de comunicação

Shannon descreveu uma arquitetura básica do processo de comunicação e qual caminho que uma determinada mensagem percorre até ser entregue, concluindo seu ciclo de vida, a Figura 3 apresenta a arquitetura básica de Shannon:

Figura 3 - "A Mathematical Theory of Communication" - Uma teoria matemática da Comunicação.



Fonte: Shannon, 1948.

Ao demostrar tal arquitetura, Shannon evidenciou que cada sistema de comunicação pode ser observado em componentes distintos e, portanto, estes são passíveis de serem tratados separadamente, cada qual com a perspectiva adequada para seu estudo e compreensão. Ficou claro que para Shannon este modelo tinha aplicações que iriam além da teoria da comunicação,

extrapolando conceitos utilizáveis para computadores, telefonia e outros campos que dependem massivamente de troca de informações.

4.3. Representação digital

Para Shannon, o conteúdo da mensagem era irrelevante para o processo de transmissão. Isto é, não importava o que a mensagem representava, somente como ela era representada para o canal, necessariamente um amontoado de zeros e uns, ou seja, uma representação digital. Esta visão para sua época pareceu pouco ortodoxa, entretanto ela unificou o formato que a engenharia das comunicações trabalhava com codificação de mensagens.

4.4. Representação da eficiência: codificação da fonte

O objetivo básico da codificação da fonte foi a remoção da redundância, ou seja, não transmitir várias vezes o mesmo fragmento de informação. Hoje este tema é investigado como compressão de dados, porém foi a partir do trabalho de Shannon que o "custo" da transferência de informações passou a ser considerado, possibilitando a otimização do processo de comunicação.

Shannon em sua obra além de propor um método eficiente de quantificação de informação, propôs um conjunto de conceitos e métodos que objetivavam o aprimoramento da maneira como são transferidas as informações, como se realiza a comunicação, como são mensurados ruídos e formatos de modo a otimizar recursos e reduzir o esforço no processo da comunicação em síntese, e assim Shannon alterou o modo como vemos a relação entre informação, ruídos e recursos.

Objetivamente, é neste processo e neste contexto que se apresenta a possibilidade de quantificar e analisar algo tão complexo quanto a informação, que embora de difícil definição, por meio da obra de Shannon, pôde ser reduzida em aspectos específicos a uma fórmula matemática aplicável.

Reestabelecendo a conexão com a definição original da entropia como medida da desordem e amparando-nos em conceitos probabilísticos, a entropia de Shannon refere-se a incerteza em uma distribuição de probabilidades.

Para Kapur e Kesavan (1992) a incerteza é de forma ampla mais complexa e não pode ser definida de forma singular; cabe porém em sua definição, três tipos de incertezas a saber:

- Incerteza determinística: em que os estados de um sistema não são determinados e portando desconhecidos;
- Incerteza entrópica: em que os estados de um sistema são conhecidos, porém é desconhecida as chances, ou probabilidade, da incidência de cada estado;
- Incerteza probabilística: em que são conhecidos os estados e a distribuição de probabilidade de incidências destes.

Entretanto não há variáveis ou evidências suficientes para determinação de qual estado irá ocorrer com certeza, uma vez que na teoria da informação a entropia se refere a uma incerteza probabilística. A definição formal da entropia pode ser observada pela fórmula abaixo:

$$H = \sum_{n=1}^{c} - \mathrm{p_n}\ log\left(\frac{1}{p_n}\right) \quad (1)$$

Vale ressaltar a definição de Shannon para a quantidade de informação de um determinado evento, considerando que a probabilidade de incidência do evento n como p(n), a quantidade de informação associada a este evento será:

$$I_n = log\left(\frac{1}{p_n}\right) \quad (2)$$

Notadamente, quanto maior for a probabilidade de ocorrência do evento x, menor será a quantidade de informação que este evento carrega em si. Shannon escolheu a função logarítmica, entre outros motivos por ser uma função crescente em relação a seu argumento, deste modo a relação da incidência de eventos e a informação nela contida é diretamente proporcional; além disso log(1) = 0; objetivamente coerente com a noção de que quanto maior a probabilidade de incidência de um determinado símbolo, menor será a quantidade de informação, neste caso, 100% de probabilidade, representa zero de informação contida. A unidade de medida para a Informação é o bit, portanto assume-se o logaritmo na base 2.

Enquanto a medida de Informação refere-se diretamente a incidência de um determinado evento, como por exemplo a transferência de um símbolo, a entropia é a quantidade de informação contida em um conjunto de eventos, neste caso, um conjunto de símbolos.

A unidade de medida da entropia é mensurada por "bits por símbolo". A entropia da informação para um conjunto de símbolos é sempre positiva, com a única exceção para se probabilidade de um dos eventos for 100% e os demais eventos possuírem probabilidades nulas.

A entropia atinge seu valor máximo quando distribuição da incidência de I valores em S símbolos for regular, ou seja, ocorre o valor máximo da entropia, quando todos os eventos são equiprováveis.

4.5. Maximização De Entropia

Considerando as propriedades da entropia da informação, a entropia de um conjunto de símbolos alcança seu ponto máximo, ou seja, entropia máxima, quando todos os seus símbolos, ou eventos possuem probabilidade de ocorrência equivalentes.

Jaynes (1957) propôs um princípio que eleva a probabilidade de ocorrência de cada símbolo amparado por restrições lineares. Assim o princípio de maximização de entropia é expressado formalmente como um problema de otimização que visa encontrar a melhor distribuição de probabilidades de um conjunto de símbolos, ou eventos, respeitando as restrições lineares atribuídas ao problema.

O princípio de maximização de entropia (JAYNES, 1957) é uma função não linear, portanto sua solução depende de algoritmos de busca iterativa e restrições lineares de modo a homogeneizar a probabilidade dos eventos e obter a entropia máxima; neste interim o processo iterativo da maximização de entropia lança mão a outros recursos de otimização, como os multiplicadores lagrangeanos (KAPUR e KESAVAN, 1992). A saber, o método dos multiplicadores de Lagrange visam identificar extremos máximos e mínimos de uma dada função, neste caso, aprimorando a homogeneidade das probabilidades de um sistema entrópico e consequentemente atingindo a máxima entropia.

Trazido ao contexto da presente pesquisa, a maximização de entropia é utilizada no processo de classificação de informações textuais, neste caso a modelagem de maximização de entropia produz distribuições de probabilidades composta de fragmentos textuais e a probabilidade de estes pertencerem ou não a uma determinada classe. Ainda, as restrições lineares são realizadas por meio da seleção de características textuais (RATNAPARKHI, 1998), que aumentam as chances de uma correta seleção de classes, dado um problema de processamento de linguagem natural.

Em síntese, o princípio de maximização de entropia é utilizado para uma implementação de aprendizado de máquina como um classificador, dotada da capacidade de

aprendizado estatístico. Seu modelo de comparação se dá por meio de um conjunto de exemplares de um determinado problema, classificado previamente, constituindo o conjunto de treinamento e, portanto, tratando-se de um modelo de aprendizado supervisionado de máquina.

Tal técnica já deu provas de sua eficiência (CARDOSO, ESTEVES e FONSECA, 2014). O princípio da entropia máxima mostrou na maior parte das vezes melhores resultados na classificação de textos. Considere C um vetor de características textuais (anotações sintáticas, dicionários e outras formas de evidenciar características relevantes) e L a classe selecionada para tais características a entropia pode ser calculada por meio do condicionamento destas restrições.

5 O MÉTODO DE INFERÊNCIA DE EMOÇÕES E SUAS FERRAMENTAS

Neste capitulo é apresentado um método para a inferência de emoções em texto por meio de um protótipo proposto; além das etapas e tarefas necessárias para o método é apresentado um conjunto de conceitos complementares, contextualizando o método e suas ferramentas.

5.1. NLTK

O *Natural Language Toolkit* (NLTK), desenvolvido pela *University of Pennsylvania* consiste de uma plataforma *open-source* desenvolvida na linguagem de programação Python® (BIRD, KLEIN e LOPER, 2009) com intuito de incentivar e facilitar os trabalhos com processamento de linguagem natural.

A plataforma conta com um extenso e completo conjunto de ferramentas, direcionado a resolução das mais variadas tarefas do processamento de linguagem natural; além destas a plataforma conta com um conjunto de corpora e recursos léxicos de diversas línguas.

5.2. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) é uma medida numérica utilizada para mensurar a importância de uma palavra em um determinado documento e se baseia em qual a frequência do uso de uma palavra em um documento e também qual sua relevância para um conjunto de documentos investigados.

$$tfidf(t, d, D) = tf(t, d)\ X\ idf(t, D) \tag{3}$$

Em que t representa os termos, ou palavras; d a representação de cada documento; D representa uma coleção de documentos (SALTON e BUCKLEY, 1998).

5.3. Ganho de informação

O ganho de informação visa o ranking das palavras por meio da redução de entropia, ou seja, quanto maior o acréscimo de informação a um determinado conjunto de palavras, menor será a quantidade de incerteza deste conjunto.

Neste contexto, uma palavra carrega muita informação, quando ela é muito relevante no texto em que ela é utilizada. Assim como na maximização de entropia este recurso usa a função de entropia de modo a medir a quantidade de entropia um determinado conjunto de treinamento S possui.

$$Entropia\ (S) = \ \sum_{n=1}^{c} - \mathrm{p_n}\ log_2\ \mathrm{p_n} \tag{4}$$

Após a determinação da medida de entropia do conjunto de treinamento é aplicado a equação de ganho de informação, conforme figura a seguir:

$$IG\ (S,t) = \ \sum \quad \left|\frac{S_v}{S}\right|\ Entropia\ (S_v) \tag{5}$$

Onde v representa o valor do termo analisado; no presente trabalho tal medida é obtida por meio da métrica TF-IDF.

5.4. O protótipo

Como um dos objetivos desta pesquisa é promover a inferência da predominância de emoções em textos provenientes de redes sociais, a escolha de tais textos acaba condicionada à complexidade de encontrar características relevantes devido ao tamanho das postagens, que possui na maioria das vezes um número limitado de palavras.

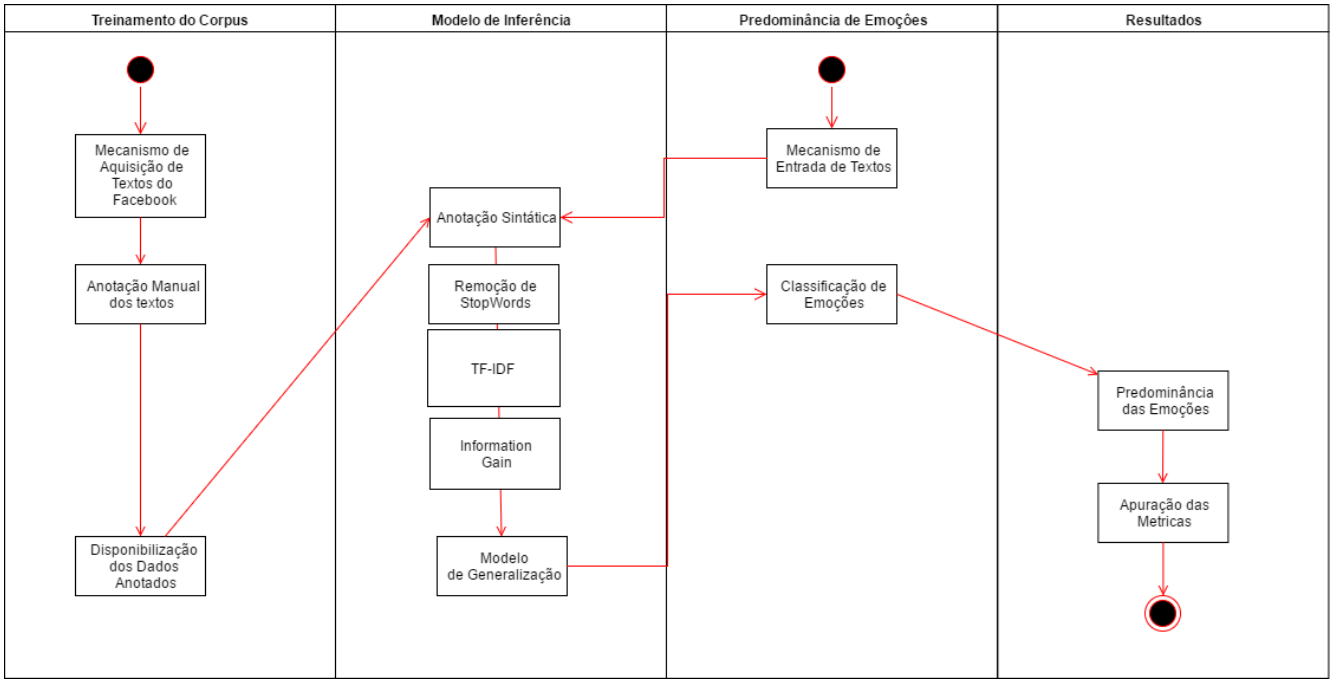
Desta forma, para a inferência das emoções em texto foi implementado um classificador baseado na teoria da informação, mais especificamente um classificador de entropia máxima. Este protótipo foi desenvolvido com a linguagem de programação Python® e a resolução do problema de linguagem natural foi desenvolvida baseada na biblioteca de código livre NLTK.

Como exemplo de texto de rede social, observa-se o fragmento a seguir: "o amor e uma fonte inesgotável de reflexão, profunda como a eternidade, alta como o céu, vasta como o universo". Tal texto foi extraído da rede social *facebook* por meio de uma ferramenta de *crawling*.

O processo de inferência de emoções em textos é uma tarefa subjetiva que carrega um alto teor de incerteza, e devido a esta característica o princípio de maximização de entropia foi escolhido para a implementação da ferramenta proposta. A essência de tal princípio e que norteou a proposição do mesmo e sua própria natureza de otimização da incerteza sugere um caminho promissor para a solução do problema proposto.

O diagrama apresentado na figura 4 demostra o mecanismo proposto para a inferência de predominância de texto.

Figura 4 - Mecanismo de Inferência de Emoções



Fonte: produção própria.

O protótipo proposto foi planejado e desenvolvido em 4 camadas, a saber:

5.4.1. Treinamento do corpus

Camada responsável pela aquisição de fragmentos textuais de redes sociais, no presente trabalho foram utilizados exclusivamente textos provenientes da rede social *facebook*[2]. O mecanismo de aquisição de textos do *facebook* foi desenvolvido por meio do módulo *BeaultifulSoup*™ da linguagem de programação Python® e seu objetivo é coletar *posts* (postagens) e comentários por meio de uma requisição ao site do *facebook* entregando as informações em uma tabela a ser trabalhada na etapa de *anotação manual dos textos*, em que as emoções predominantes foram selecionadas manualmente a fim de produzir o conjunto de treinamento a ser utilizado posteriormente pelo modelo de generalização.

5.4.2. Modelo de inferência

Esta etapa é aquela na qual se dá o tratamento e a identificação das características a serem analisadas no texto analisado. Tanto o corpus anotado (da etapa de treinamento), quanto um texto candidato as inferências de emoções passam por este processo, a fim de serem usados no modelo de maximização de entropia, deste modo a distribuição de probabilidade (p) do corpus anotado, ou conjunto de treinamento é comparada com a distribuição de probabilidade (p') de um candidato a classificação.

A primeira tarefa nesta etapa consiste na anotação sintática, ou seja, o texto processado, tem suas palavras etiquetadas conforme seu funcionamento e relacionamentos dentro das sentenças, em seguida a etapa filtra o texto e remove palavras com alta incidência e baixa relevância do corpus, a chamada remoção de *stopwords***.**

Posteriormente o texto é analisado de acordo com a relevância de cada palavra, considerando a relevância no contexto de todo o corpus, utilizando o método matemático TF-IDF, atribuindo a elas um peso que será utilizado complementarmente na tarefa de *infomation gain*, método de otimização baseado na teoria da informação e utilizado na seleção de palavras e termos com características relevantes no processo de classificação.

Por último, todas as características compiladas ao longo das tarefas anteriores são utilizadas para a produção do modelo de generalização por meio dos critérios estatísticos da maximização de entropia.

[2] Em trabalhos futuros, intentamos trabalhar com um *pool* heterogêneo de redes sociais, aprimorando o processo de seleção de características e consequentemente melhorando o modelo aqui proposto

5.4.3. Predominância de emoções

Nesta etapa ocorre a efetiva utilização do modelo e a promoção da comparação das distribuições de probabilidade, conforme proposto no modelo de maximização de entropia. O processo de classificação produz uma lista de probabilidades relacionadas a classes candidatas a seleção, tais classes são representadas, pelas emoções primárias: alegria, medo, desgosto, tristeza, surpresa e raiva.

O objetivo central do protótipo proposto é a inferência de emoções a partir de um método de aprendizado de máquina, as etapas e tarefas anteriores funcionam para servir este proposito ao criar um modelo de generalização e disponibilizar um formato ao qual por meio do mecanismo de entrada de textos**,** um fragmento textual hipotético possa ser processado no protótipo e na etapa de predominância de emoções, ser analisado conforme suas características e utilizar o modelo de generalização produzindo a classificação de emoções, que além da seleção de uma emoção candidata serve um conjunto de probabilidades das demais emoções.

5.4.4. Resultados

A etapa de resultados é quando são selecionadas as classes ou neste caso as emoções com maior probabilidade de ocorrência dadas as características textuais presentes no fragmento testado. Deste modo, determina-se qual emoção é predominante no referido fragmento.

A inferência da predominância de emoções é obtida por meio da composição das probabilidades de cada emoção estar contida no proposto fragmento textual hipotético.

Complementarmente ao processo de treinamento, classificação e inferência a etapa de resultados conta com mecanismos de apuração das métricas de eficiência do modelo dado as classes, ou neste caso as emoções primarias, isto é, mede a eficiência do treinamento, usado na composição do modelo de generalização, considerando as características descobertas ao longo do processo de utilização do protótipo.

5.5. Normalização de variações linguísticas

Reconhecer variações da linguagem em textos permite, por exemplo, maior controle do vocabulário e da correta relevância de termos em um dado *corpus*. Basicamente, a normalização pode ser dividida em morfológica, sintática e léxico-semântica

Morfológica, quando há redução das palavras, ao seu estado raiz (radical), prática conhecida como *stemming*, ou por meio da transformação de palavras a sua forma canônica, por exemplo a substituição de um verbo por sua forma no infinitivo, prática conhecida como lematização

A normalização sintática ocorre quando frases semanticamente equivalentes, mas distintas sintaticamente são traduzidas de forma única; enquanto a normalização léxico-semântica ocorre quando grupamentos semânticos (sinonímia, hiponímia e metonímia) são representados uniformemente, apresentando-se em um único conceito.

5.6. Métricas de apuração

Os experimentos do presente trabalho foram realizados com o uso da técnica de K-*folds cross-validation*, onde a amostra original é dividida aleatoriamente em k subconjuntos e desses subconjuntos, uma porção é utilizada para treinamento e outra porção é utilizada para o teste de desempenho (KOHAVI, 1997). Foram utilizadas como métricas a precisão, a cobertura e a medida-f para a análise da qualidade do modelo gerado, abaixo uma breve descrição de tais métricas.

a) A acurácia é a medida mais simples de qualquer experimento; ela demostra a razão entre a quantidade de acertos e o total de observações de um experimento, expresso como uma porcentagem, como pode ser observado na equação 5.

$$A = \frac{\text{Quantidade de Acertos}}{\text{Total de Observações}} \tag{6}$$

b) Precisão que se refere a quantidade de candidatos selecionados corretamente, dentro de um conjunto de candidatos em uma observação, conforme a equação 6.

$$P = \frac{Número\ de\ Elementos\ Relevantes\ Recuperados}{\text{Total de Elementos Relevantes Recuperados}} \tag{7}$$

c) Cobertura que se refere a fração de candidatos corretos que foram selecionados; conforme equação 7.

$$C = \frac{Número\ de\ Elementos\ Relevantes\ Recuperados}{\text{Total de Elementos Relevantes}} \tag{8}$$

d) Medida-F é a média harmônica ponderada da precisão e cobertura, conforme apresentado na equação 8. (NADEAU e SEKINE, 2007).

$$F1 = \frac{2 \text{ X P X C}}{\text{P+C}} \quad (9)$$

5.7. O *corpus* da pesquisa

O *corpus* da pesquisa é proveniente de rede social *Facebook*; foram capturados um conjunto de 5000 posts e comentários, visando a identificação de textos que se enquadrassem nas emoções: alegria, raiva, surpresa, medo, desgosto e tristeza.

Deste conjunto foram selecionados 240 textos, para cada emoção, perfazendo um corpus de 1440 documentos selecionados manualmente. Este corpus é um recurso de extrema importância para o presente trabalho, uma vez que é por meio dele que ocorre o processo de aprendizado, possibilitando a inferência de emoções em textos distintos.

Dos 240 documentos selecionados para cada emoção foi utilizado 70% do conjunto para compor os dados de treinamento e 30% foram utilizados para a composição do conjunto de testes. Abaixo um exemplo da composição do conjunto de treinamento:

Quadro 1 - Composição de Conjunto de Treinamento

Emoção	***Fragmento de Texto***
Alegria	Não é difícil estar alegre seu lado; você me motiva e me inspira, a cada dia sou mais feliz com você
Desgosto	Desgosto perdi o meu tempo deixando nosso amor ao relento. Lágrimas rolavam em meu rosto pois ao meu erro relembrar aquele desgosto. Mais uma chance eu desejo para que dessa vez eu não perca mais o seu beijo. Para mim és importante então por favor não ficai mais distante.

Medo	O medo faz parte da vida da gente. Algumas pessoas não sabem como enfrentá-lo, outras - acho que estou entre elas - aprendem a conviver com ele e o encaram não como uma coisa negativa, mas como um sentimento de autopreservação.
Raiva	Tenho raiva de mim, muito ódio de mim mesma, você sabe o porquê? porque eu não consigo parar de pensar em você! eu te odeio mais do que me odeio por que ao mesmo tempo que é tão imperfeito é muito perfeito.
Surpresa	O bom de não parar planejando o futuro, é se divertir mais com cada inédita surpresa
Tristeza	Os bons vi sempre passar no mundo graves tormentos; e para mais me espantar os maus vi sempre nadar em mar de contentamentos.

Fonte: produção própria.

5.8. Execução

Considerando o fluxo do protótipo proposto, após o processo de aquisição de textos provenientes ao *facebook,* os textos são processados a fim de se obter características relevantes, para que o algoritmo de aprendizagem aprenda a quais classes determinados textos pertencem, produzindo a inferência de emoções; nesta sessão serão apresentados exemplos de textos, por emoções para cada etapa de processamento para a aquisição de características.

A anotação sintática atribui além das palavras como características o papel sintático de determinadas palavras. Foi utilizado o anotador sintático da plataforma NLTK, treinado por meio do corpus (ALUÍSIO et al., 2003); a presente pesquisa não considerou a assertividade do anotador sintático. A figura abaixo ilustra fragmentos de textos anotados, conforme sua estrutura sintática:

Quadro 2 - Palavras Anotadas

Alegria	Desgosto	Medo	Raiva	Surpresa	Tristeza
Não - ADV	Desgosto - NOUN	O - DET	Tenho - VERB	O - DET	Os - DET
é - VERB	perdi - NOUN	medo - NOUN	raiva - NOUN	bom - ADJ	bons - ADJ
difíl - NOUN	o - DET	faz - VERB	de - ADP	de - ADP	vi - VERB
estar - VERB	meu - PRON	parte - NOUN	mim - PRON	não - ADV	sempre - ADV
alegre - ADJ	tempo - NOUN	da - ADP	muito - ADV	parar - VERB	passar - VERB
ao - ADP	deixando - VERB	vida - NOUN	ódio - NOUN	planejando - ADP	no - ADP
seu - PRON	nosso - PRON	da - ADP	de - ADP	o - DET	mundo - NOUN
lado - NOUN	amor - NOUN	gente - NOUN	mim - PRON	futuro - NOUN	graves - ADJ
você - PRON	ao - ADP	Algumas - PRON	mesma - PRON	é - VERB	tormentos - NOUN
me - PRON	relento - NOUN	pessoas - NOUN	você - PRON	se - CONJ	e - CONJ
motiva - VERB	Lágrimas - NOUN	não - ADV	sabe - VERB	divertir - VERB	para - ADP
e - CONJ	rolavam - VERB	sabem - VERB	o - DET	mais - ADV	mais - ADV
me - PRON	em - ADP	como - ADV	porquê - ADV	com - ADP	me - PRON
inspira - NOUN	meu - PRON	enfrentá-lo - NOUN	porque - CONJ	cada - PRON	espantar - VERB
a - DET	rosto - NOUN	outras - PRON	eu - PRON	inédita - ADJ	os - DET
cada - PRON	pois - CONJ	- - .	não - ADV	surpresa - NOUN	maus - ADJ
dia - NOUN	ao - ADP	acho - VERB	consigo - VERB		vi - VERB
sou - VERB	meu - PRON	que - PRON	parar - VERB		sempre - ADV
mais - ADV	erro - NOUN	estou - VERB	de - ADP		nadar - VERB
feliz - ADJ	relembrar - VERB	entre - ADP	pensar - VERB		em - ADP
com - ADP	aquele - PRON	elas - PRON	em - ADP		mar - NOUN
você - PRON	desgosto - NOUN	aprendem - VERB	você - PRON		de - ADP

Fonte: produção própria.

Após esta etapa, o protótipo faz a limpeza de termos com baixa relevância, o chamado processo de remoção de *stopwords*, bem como o uso do algoritmo de *stemming* da plataforma NLTK. Abaixo um exemplo de fragmentos de textos processados:

Quadro 3 - Textos processados por stopwords

Emoção	***Fragmento de Texto***
Alegria	não é difícil estar alegr lado; motiv inspira, cad dia feliz
Desgosto	desgost perd temp deix amor relento. lagrim rol rost pois erro relembr desgosto. chanc desej dess vez perc beijo.para mim és importante então por favor não ficai mais distante..
Medo	med faz part vid gente. algum pesso sab enfrentá-lo, outr - acho - aprend conviv encar cois negativa, sentiment
Raiva	raiv mim, ódi mim mesma, sab porque? porqu consig par pens você! odei odei temp é tã imperfeit é perfeito.
Surpresa	bom par planej futuro, é divert cad inédit surpres
Tristeza	bons vi sempr pass mund grav tormentos; espant maus vi sempr nad mar contentamentos.

Fonte: produção própria.

Em seguida é realizado o processo de análise de relevância de todos os termos processados, considerando todos os textos para cada emoção; o protótipo utiliza os pesos baseado na métrica de TF-IDF e na métrica de ganho de informação, basicamente para selecionar o quão relevante é um termo em relação ao conjunto de treinamento.

Trabalhar com a tarefa de classificação baseado em raciocínio supervisionado de máquina se baseia fortemente em quais características textuais serão usadas. As tarefas de TF-IDF e ganho de informação são utilizadas, para selecionar os fragmentos mais relevantes do conjunto de treinamento onde

TDF-IDF apresenta uma medida da raridade de termos e ganho de informação uma medida mais elaborada de o quão comum é um termo em uma classe particular, neste caso, quão comum é um termo em determinada emoção para a garantia da correta inferência de predominância; e também o quão comum é este termo para todas as classes.

A tabela 4 a seguir, mostra um exemplo de palavras com alta relevância para cada classe referida.

Quadro 4 - Relevância de Palavras

Alegria	Raiva	Tristeza	Desgosto	Medo	Surpresa
feliz	cólera	infeliz	nojo	obscuro	surpreendente
vivo	furor	depressão	asco	vil	milagre
entusiasmo	agressivo	ruim	maldoso	temor	pasmo
gosto	antipatia	servil	nausea	susto	misterio
prazer	odioso	trevas	repulsa	ruim	estupefato

Fonte: produção própria.

6 RESULTADOS OBTIDOS

Aqui são apresentados e discutidos os resultados obtidos em decorrência da implementação do protótipo proposto, bem como uma comparação dos resultados obtidos considerando o método de maximização de entropia e o método bayesiano.

Os testes realizados no protótipo proposto no presente trabalho tiveram por objetivo estabelecer a inferência de predominância de emoções em um corpus baseado em 1440 textos provenientes de posts e comentários do *facebook*; estes textos são compostos de 240 exemplares para cada emoção, considerando as emoções primárias propostas por Damásio (2000). Além do referido corpus, utilizou-se uma lista de sinônimos, para as palavras: alegria, tristeza, raiva, desgosto, surpresa e medo.

Todos os textos capturados do *Facebook*, foram persistidos em arquivos de texto simples e todas as matrizes de características para o corpus foi trabalhada diretamente na memória do computador, em decorrência da grande quantidade de características geradas.

Considerando o modelo de validação cruzada (*cross-validation*) onde uma porção do *corpus* (30%) foi utilizada como conjunto de testes e o restante utilizado para compor o modelo de generalização; o protótipo obteve uma assertividade média na ordem de 90% na inferência de emoções, a tabela abaixo compõe os índices de avaliação, considerando todas as emoções.

Tabela 1 - Medidas de qualidade por emoção (Bayes)

Emoção	***Precisão***	***Cobertura***	***Medida-F***
Alegria	0,756	0,903	0,823
Desgosto	0,980	0,861	0,925
Medo	0,885	0,958	0,920
Raiva	0,969	0,875	0,920
Surpresa	0,960	0,972	0,986
Tristeza	0,859	0,847	0,853

Fonte: produção própria.

O resultado foi superior, quando comparado com um modelo baseado no método de Naive Bayes, que obteve taxa média de assertividade de 88%, considerando as mesmas

características selecionadas para o experimento com maximização de entropia, como pode ser observado na tabela abaixo:

Tabela 2 - Medidas de qualidade por emoção (Maxent)

Emoção	***Precisão***	***Cobertura***	***Medida-F***
Alegria	0,930	0,736	0,822
Desgosto	0,955	0,875	0,913
Medo	0,925	0,861	0,892
Raiva	0,852	0,958	0,902
Surpresa	0,899	0,986	0,940
Tristeza	0,780	0,889	0,831

Fonte: produção própria.

O próximo estágio do experimento é baseado no uso do modelo de generalização em um novo texto coletado do *facebook*, que não faça parte do conjunto de treinamento, nem no conjunto de testes; o processo de inferência de emoções se dá por meio dos pesos encontrados no texto a ser classificado em relação a distribuição de probabilidade de pesos produzida no modelo obtido do conjunto de treinamento. A figura abaixo apresenta uma lista de textos do conjunto de treinamento e os pesos atribuídos, para cada emoção presente em sua composição.

Quadro 5 – Peso Por Emoção

Texto	Alegria	Desgosto	Medo	Raiva	Surpresa	Tristeza
o medo faz parte da vida da gente. algumas pessoas não sabem como enfrentá-lo, outras - acho que estou entre elas - aprendem a conviver com ele e o encaram não como uma coisa negativa, mas como um sentimento de autopreservação	5.341	6.601	20.316	4.016		
o medo de perder tira a vontade de ganhar.	5.006	3.074	21.523			3.819
que o outro saiba quando estou com medo, e me tome nos braços sem fazer perguntas demais. que o outro note quando preciso de silêncio e não vá embora batendo a porta, mas entenda que não o amarei menos porque estou quieta.	7.490	1.471	19.104			7.392
a coragem alimenta as guerras, mas é o medo que as faz nascer.	2.651	3.973	21.816			7.554
o medo de ser pego é melhor do que a sensação de fazer o errado, mas os dois juntos é a união perfeita.	4.340		20.841	4.312		3.163

Fonte: produção própria.

A utilização do modelo de generalização, foi realizada por meio da submissão de um texto ao um webservice, configurado no protótipo, visando facilitar o processo de inferência de emoções; foi selecionado um exemplo para cada emoção, considerando a seleção manual de sua emoção predominante e sua classe manual foi comparada ao resultado obtido do protótipo, conforme a tabela abaixo.

Quadro 6 - Predominância de emoção em texto aleatório

Texto	Alegria	Desgosto	Medo	Raiva	Surpresa	Tristeza	Seleção Manual
Ser grato é um excelente remédio para manter o equilíbrio mental, saiba o por quê!autopreservação	**10.591**		7.649		6.356	8.255	Alegria
tem pessoas que são tão baixas que eu tenho nojo só de ouvir seus nomes.	8.546	**18.590**			6.196	7.284	Desgosto
o valente tem medo do seu adversário; o covarde tem medo do seu próprio temor.	2.625		**20.372**		3.181	4.682	Medo
engulo o choro, e transformo a tristeza em raiva. os sentimentos são meus e eu faço o que quiser com eles.	6.930			**13.361**	6.233	7.652	Raiva
o amor? ah, o amor é aquele malandro que chega no sapatinho e pega a gente de surpresa. é um moleque travesso que adora brincar com a gente.	12.922		4.912		**14.085**	4.101	Surpresa
Fico triste quando alguem me ofende, mas, com certeza, eu ficaria mais triste se fosse eu o ofensor... magoar alguem é terrivel - Chico Xavier			2.328	4.732	1.288	**20.058**	Tristeza

Fonte: produção própria.

7 CONCLUSÃO E PERSPECTIVAS

O presente trabalho objetivou o desenvolvimento de um protótipo voltado a inferência de emoções primárias, conforme proposto por Ekman e Friesen (1978) e defendido por Damásio (2000), por meio do método estatístico de maximização de entropia.

A tarefa de inferência de emoções possui grande incerteza, uma vez que as emoções são baseadas em critérios subjetivos. A parte mais importante e significante da pesquisa foi baseado na revisão bibliográfica acerca das emoções, sentimentos, incerteza e informação, visando um bom repertório de critérios visando direcionar o processo da pesquisa. Para o protótipo foi escolhido a linguagem de programação Python® e a plataforma de processamento de linguagem natural NLTK, visando a velocidade de processamento e obtenção de resultados, quando comparado com outros métodos durante o processo de pesquisa e construção do protótipo.

Após a conclusão da implementação do método, iniciou-se a etapa de testes e experimentações a fim de validar o presente trabalho baseado na hipótese de inferência de emoções em textos do *Facebook*. O corpus de treinamento, bem como os textos de testes foram obtidos por meio de consultas aleatórias a páginas do *Facebook*. O processo de etiquetagem dos textos segundo suas emoções, foi baseado em seleção manual, baseado na percepção do autor da presente pesquisa. A avaliação da qualidade do modelo se deu por meio do processo de validação Cruzada (*cross-validation*), obtendo a taxa de assertividade média de 90% entre as classes/emoções estudadas.

Os experimentos realizados, comprovaram a viabilidade de inferência de emoções baseado em um modelo de aprendizagem de máquina por meio de um mecanismo de raciocínio estatístico. O protótipo obteve resultados satisfatórios e esperados, corroborando as hipóteses que fomentaram os objetivos deste trabalho, quando comparado a outros trabalhos similares, como pode ser observado na tabela a seguir.

Tabela 3 - Resultados de extração de emoções

Emoção	Nº. de textos	Nº. de acertos	Acurácia
Alegria	116	69	59%
Desgosto	78	60	77%
Medo	20	16	80%
Raiva	18	9	50%
Surpresa	7	6	86%
Tristeza	63	45	71%

Fonte: (DOSCIATTI, et al., 2012).

Apesar da eficiência do método de maximização de entropia, a curta extensão de textos provenientes da rede social *Facebook*, colabora com o super-ajuste (*overfitting*) do modelo (PAPES, PETERSON e SOBERON, 2007).

Espera-se para trabalhos futuros a coleta de um corpus heterogêneo, baseado em múltiplos tipos de texto, bem como o uso de técnicas de seleção de características que demandem menos supervisão humana.

A emoção é baseada em percepção pessoal, logo, vale também para uma próxima pesquisa a investigação do uso de um modelo conjunto, baseado em características de textos independentes de seu autor e características especificas, elaboradas por meio de textos específicos para cada autor, ou conjuntos de indivíduos, visando o grupamento de percepções, baseado em critérios a serem analisados futuramente. Objetiva-se também o aprimoramento do modelo de maximização de entropia, por meio do uso de Redes Neurais Recorrentes - RNR, dando a pesquisa mecanismos mais escaláveis em relação ao presente trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

AFTAB, O.; CHEUNG, P.; KIM, A.; THAKKAR, S. e YEDDANAPUDI, N. **Information Theory after Shannon's 1948 Work**. Project History, Massachusetts Institute Of Tecnhnology, 2001; http://mit.edu/6.933/www/Fall2001/Shannon2.pdf, 2005.

ALUÍSIO, S., PELIZZONI, J., MARCHI, A.R., de OLIVEIRA, L., MANENTI, R., MARQUIAFÁVEL, V. 2003. **An account of the challenge of tagging a reference corpus for brazilian portuguese**. In: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language. PROPOR 2003.

BIRD. S.; KLEIN E.; LOPER. E.; **Natural Language Processing With Python, Analyzing Text with the Natural Language ToolKit**. O'Reilly Media, Beijing. 2009.

CARDOSO, J. M. M.; ESTEVES, M. R. e FONSECA, C. M. M. **SoFly: meet the social engineering**. Faculdade De Ciências E Tecnologia - Universidade De Coimbra. 2014.

CARVALHO, W. S. **Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina**. Dissertação (Mestrado em Ciências da Computação) - Universidade de São Paulo, São Paulo, 2012.

CASTELLS, M. **The Impact of the Internet on Society A Global Perspective**. University of Southern Califórnia, 2014.

CHAUDIRON, S. **Technologies linguistiques et. modes de représentation de l'information textuelle**. Documentaliste – sciences de l'information, v. 44, n. 1, p. 30–39, 2007.

CHERPAS, C. **Natural language processing, pragmatics, and verbal behavior**. Anal Verbal Behav. 10:135-47, 1992.

CHIAVENATO, I. **Introdução à teoria Geral da Administração. 6. ed**. Rio de Janeiro: Campus, 2000.

CISCO. **Cisco Visual Networking Index: Forecat and Methodology, 2014-2019**. Disponível em: <http://s2.q4cdn.com/230918913/files/doc_downloads/report_2014/white_paper_c11-481360.pdf>. Acesso: 05 dez. 2016.

CUCS, CORNELL UNIVERSITY (CU) COMPUTER SCIENCE (CS). **Sentiment Analysis**. CS 40th anniversary Symposium, p. 26-27, 2005.

DALE, R. **Classical approaches to natural language processing**. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) Handbook of natural language processing. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

DAMASIO, A. **Descartes error: emotion, reason, and the human brain**. New York: Quill, 2000.

DARWIN, C. **The Expression of the Emotions in Man and Animals**. Chicago: University of Chicago Press, 1872.

DOSCIATTI, M. M.; MARTINAZZO, B.; PARAISO, E. C. **Identifying Emotions in Short Texts for Brazilian Portuguese**. In: IV International Workshop on Web and Text Intelligence, Curitiba, 2012.

EKMAN, P.; FRIESEN, W.V. **Constants across cultures in the face and emotion**. Journal of Personality and Social Psychology, Vol 17(2), 1971.

EPSTEIN, I. **Teoria da Informação**. 2a Ed. São Paulo: Ática, 1988.

FISCHER, E. **A necessidade da arte**. 4. Editora Rio de Janeiro: Zahar Editores, 1973.

GOLEMAN, D. **Inteligência emocional**. Portugal: Temas e Debates. 1997.

GOOD, I. J. **Speculations Concerning the First Ultraintelligent Machine**. Advances in Computers. vol. 6. Archived May 1, 2012 at the Wayback Machine, 1965.

HIPPISLEY, A. **Handbook of Natural Language Processing, Second Edition**. University of Kentucky. Nitin Indurkhya & Fred J. Damerau (Eds.), p. 31-58, 2010.

JAYNES. E. T., **Information Theory and Statistical Mechanics**. Physical Review. vol. 106, no. 4, pp. 620-630; May 15, 1957.

JURAFSKY, D.; MARTIN, J. H. **Speech And Language Processing**. Englewood Cliffs, NJ, USA: Prentice Hall, 2008.

KAPUR, J.N. e KESAVAN, H. K. **Entropy Optimization Principles With Application**. Academic Press, Inc, 1992.

KOBASHI, N. Y.; TÁLAMO, M. F. G. M. **Informação: fenômeno e objeto de estudo da sociedade contemporânea**. Transinformação, Campinas, v. 15, p. 7-22, 2003.

KOCH, I. V.; TRAVAGLIA, L. C. **Texto e coerência**. 7. ed. São Paulo: Cortez, 1997.

KOHAVI, Ron. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 1137–1143, 1997.

KONKIEWITZ, C. E. **Tópicos em Neurociência Clínica**. Editora UFGD, 2009.

LJUNGLÖF, P.; WIRÉN, M. Syntactic parsing. In: INDURKHYA, N.; DAMERAU, F. J. **Handbook of Natural Language Processing**. 2. ed. Boca Raton, FL: Chapman & Hall/CRC, p. 59-91, 2010.

MACLEAN, P. D. **The triune brain in evolution: role in paleocerebral functions**. New York: Plenum Press. 1990.

MATTOS, R. S.; VEIGA. A. **Otimização de Entropia: Implementação Computacional dos Princípios MaxEnt e MinxEnt**. Revista Pesquisa Operacional, 2002.

MOENS, M. F.; UYTTENDAELE, C.; DUMORTIER, J. **Information extraction from legal texts: the potential of discourse analysis**. International journal of human–computer studies, v. 51, n. 6, p. 1155–1171, 1999.

NADEAU, D.; SEKINE. S. **A Survey of named entity recognition and classification**. Em Linguisticae Investigationes, páginas 3-26, Janeiro 2007.

NORVIG, P.; RUSSELL, S. J. **Artificial Intelligence – A Modern Approach. 2. ed**. Englewood Cliffs, NJ, USA: Prentice Hall, 2003.

ORTONY, A., e TURNER, T. J. **What's basic about basic emotions?** Psychological Review. 97, 315-331, 1990.

PAK, A. e PAROUBEK, P. **Twitter as a corpus for sentiment analysis and opinion mining**. In Proceedings of LREC, 2010.

PALMER, D. D. **Text preprocessing**. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.) Handbook of natural language processing. 2. ed. Boca Raton: Chapman & Hall/CRC, 2010.

PAPES, J. W. **A Proposed Mechanism Of Emotion**. Arch NeurPsych. 1937.

PAPES. M.; PETERSON, A. T.; SOBERON., J. **Rethinking receiver operating characteristic analysis applications in ecological niche modeling**. Natural History Museum and Biodiversity Research Center, the University of Kansas, Lawrence, KS 66045 USA. 2007.

PIAGET, J. **The relation of affetivity to intelligence in the mental development of the child**. [transl. by Pitsa Hartocollis]. In Bulletin of the Menninger Clinic. – 1962, vol. 26, no 3. Three lectures presented as a series to the Menninger school of psychiatry March, 6, 13 and 22, 1961. Publicação original em língua inglesa, 1962.

PICARD, R. **Affective Computing**. The MIT Press. 1997.

PINTO, A. C. **Psicologia Geral**. Lisboa: Universidade Aberta. No 227. (340 páginas). ISBN: 972-674-339-7. DL: 164485/01, 2001.

PLUTCHIK, R. **Emotion, a psychoevolutionary synthesis**. New York: Harper & Row, 1980.

POTTIER, B. **A substância do significado**. In: Lingüística geral: teoria e descrição. trad. adap. Walmírio Macedo. Rio de Janeiro: Presença; Universidade Santa Úrsula. p. 61–96. 1978.

RATNAPARKHI, A. **Maximum Entropy Models for Natural Language Ambiguity Resolution**, 1998.

RÜDIGER, F. **Introdução à Teoria da Comunicação**. São Paulo: Edicon, 1998.

RUMELHART, D. E.; SMOLENSKY, P; McCLELLAND, J.; HINTON, G. E. **Schemata and Sequential thought processes in PDP models**. In: McCLELLAND, J.; RUMELHART, D. E. (Ed.). Parallel Distributed Processing, v. 2. Cambridge: MIT Press, 1986.

SALTON, G.; BUCKLEY, C. **Term-weighting approaches in automatic text retrieval**. Information Processing and Management – Cornell University, Ithaca. 1998.

SARDINHA, T. B. **Linguística de Corpus**. São Paulo: Manole, v. 1. 410p., 2004.

SÉNECA, L. A. Cartas a Lucílio. Madrid: Fundação Calouste Gulbenkian, 1991.

TAILLE, Y. et al. **Piaget, Vygotsky e Wallon: teorias psicogenéticas em discussão**. São Paulo, Summus, 1993.

WEISS, S.; KULIKOWSKI, C. **Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems**. San Mateo, CA, 2001.