UNIVERSIDADE FEDERAL TECNOLÓGICA FEDERAL DO PARANÁ
ESPECIALIZAÇÃO EM BANCO DE DADOS

LILIAN DO NASCIMENTO ARAUJO LAZZARIN

TÉCNICA PARA MINERAÇÃO DE TEXTOS NA ANÁLISE DE SENTIMENTOS: Um estudo de caso em uma Instituição de Ensino Técnico, Tecnológico e Superior

MONOGRAFIA DE ESPECIALIZAÇÃO

PATO BRANCO
2017

LILIAN DO NASCIMENTO ARAUJO LAZZARIN

TÉCNICA PARA MINERAÇÃO DE TEXTOS NA ANÁLISE DE SENTIMENTOS: Um estudo de caso em uma Instituição de Ensino Técnico, Tecnológico e Superior

Monografia apresentada ao II Curso de Especialização em Banco de Dados, da Universidade Tecnológica Federal do Paraná, Campus Pato Branco, como requisito para obtenção do título de Especialista.

Orientador: Prof. Dr. Richardson Ribeiro.

PATO BRANCO

2017

TERMO DE APROVAÇÃO

TÉCNICA PARA MINERAÇÃO DE TEXTOS NA ANÁLISE DE SENTIMENTOS: Um Estudo de Caso em uma Instituição de Ensino Técnico, Tecnológico e Superior.

por

LILIAN DO NASCIMENTO ARAUJO LAZZARIN

Este Trabalho de Conclusão de Curso foi apresentado em 24 fevereiro de 2017 como requisito parcial para a obtenção do título de Especialista em Banco de Dados. O(a) candidato(a) foi arguido(a) pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Richardson Ribeiro
Prof.(a) Orientador(a)

_____
Mariza Miola Dosciatti
Membro titular

_____
Dalcimar Casanova
Membro titular

"O Termo de Aprovação assinado encontra-se na Coordenação do Curso"

RESUMO

LAZZARIN, Lilian N. A..TÉCNICA PARA MINERAÇÃO DE TEXTOS NA ANÁLISE DE SENTIMENTOS: Um estudo de caso em uma Instituição de Ensino Técnico, Tecnológico e Superior. 2017. Monografia (II Curso de Especialização em Banco de Dados) – Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

Neste trabalho é proposta a aplicação de técnicas de mineração de dados em textos extraídos da rede social Twitter para análise de sentimentos. Empresas tem se beneficiado dessa ferramenta com a geração de informações sobre seus produtos e/ou serviços, bem como observar o comportamento dos usuários que podem auxiliá-los no direcionamento de promoções, lançamentos, marketing, etc. Com a crescente demanda pelo uso das redes sociais bem como o uso de textos não estruturados, a interpretação das avaliações dos usuários pelas empresas ou interessados tem se tornado uma tarefa difícil. Neste trabalho, a partir de dados extraídos e pré-processados de uma rede social, pretende-se aplicar abordagens de mineração de texto para interpretação dos dados no que se refere ao nível emocional dos usuários de uma determinada instituição de ensino. Espera-se que, a identificação do estado emocional dos indivíduos que usam esta rede social pode ajudar a identificar o nível de satisfação desses indivíduos quanto a assuntos relacionados a instituição, e também direcionar esforços/melhorias para assuntos onde foram identificadas emoções como tristeza, raiva ou decepção em comentários, e com isso sanar problemas até então desconhecidos ou que passavam desapercebidos pela instituição de ensino.

Palavra-chave: análise de sentimentos, rede social, mineração de textos

ABSTRACT

LAZZARIN, Lilian N. A. Approach for text mining in feeling analysis: A case study in a Technical, Technological and Superior Institution. 2017. Monography (II Specialization Course in Database) - Federal University of Technology - Parana. Pato Branco, 2017.

This work proposes the application of data mining techniques in texts extracted from the social network Twitter for analysis of feelings. Companies have benefited from this tool with the generation of information about their products and / or services, as well as observing the behavior of users that can assist them in targeting promotions, launches, marketing, etc. With the increasing demand for the use of social networks as well as the use of unstructured texts, the interpretation of user evaluations by companies or stakeholders has become a difficult task. In this work, from data extracted and pre-processed from a social network, we intend to apply text mining approaches for data interpretation regarding the emotional level of users of a given educational institution. It is expected that the identification of the emotional state of individuals using this social network can help to identify the level of satisfaction of these individuals with regard to institution-related issues and also direct efforts / improvements to issues where emotions such as sadness, anger or disappointment in comments, and thereby remedy problems unknown or unnoticed by the educational institution.

Keywords: feelings analysis;. Social network. Text mining.

LISTAS DE ABREVIATURAS E SIGLAS

ANET	-	Affective Norms for English Text
ANEW	-	Affective Norms of English Words
ANEW-BR	-	Brazilian Norms for the Affective Noms for English Wods
API	-	Application Programming Interface
ARFF	-	Attribute-Relation Format
ASCII	-	American Standard Code for Information Interchange
CBI	-	Classificador Bayesiano Ingenuo
CSV	-	Comma-Separated Values
GNU	-	GNUS's Not Unix
IADS	-	International Affective Digital Sounds
IAPS	-	International Affective Picture System
JSON	-	JavaScript Object Notation
KDD	-	Knowledge Discovery in Databases
KDT	-	Knowledge Discovery in Texts
KNN	-	K-Nearest Neighbor
MSE	-	Mean Squared Error
NL	-	Natural Language
NLTK	-	Natural Language Toolkit
OPENNLP	-	The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text
PLN	-	Processamento de Linguagem Natural
SVM	-	Support Vector Machine
WEKA	-	Waikato Environment for Knowledge Analysis

LISTAS DE FIGURAS E QUADROS

SUMÁRIO

1 INTRODUÇÃO

Uma rede social pode ser definida como uma estrutura social composta por vários indivíduos ou organizações, e os atores (indivíduos) são os nodes (nós) que estão ligados a vários outros nodes com um ou mais tipos específicos de relacionamentos como amizade, parentesco, crenças, interesses em comum entre outros (WILLIAMS, 2008). A análise de uma rede social considera as relações sociais em termos da teoria como uma rede composta por nós e laços denominada conexões. Para vários campos acadêmicos as redes sociais demonstraram como operam de forma significativa desempenhando um papel crítico na resolução de problemas, e como os indivíduos alcançam objetivos esperados nos mais difentes níveis, desde o familiar até o nível de nações (WASSERMAN e FAUST, 2011).

As redes sociais tem despertado o interesse de vários pesquisadores, devido a quantidade de informações geradas pelo seu uso frequente. [1]Estima-se que no Brasil a rede social Twitter, recebe mais de 10 mil tuítes por minuto postados diariamante chegando em um ano a 550 milhões de mensagens. [2]Em 2015 a rede social Twitter já foi mais atraente, mas ainda possui um número significativo de usuários ativos com 260 milhões, ainda é uma opção interessante para empresas divulgarem suas marcas e que o mesmo registra aproximadamente 500 milhões de tuítes todos os dias. Para as empresas, as redes sociais possibilitam medir as informações postadas para tomada de decisão, e definir métricas sobre a qualidade de seus produtos e serviços, além de conteúdos estratégicos potencializando a divulgação e propagação de suas informações.

Outro viés, é a partir de informação textual aplicar métodos que identifiquem a emoção do indivíduo com a prática da análise de sentimentos que, vem sendo pesquisado com diversos métodos nas redes sociais.

Para o contexto político (CRUVINEL), coletou dados da rede social Twitter, utilizando como expressão de busca o nome da ex-presidente do Brasil “Dilma”, os tweets extraídos foram manualmente classificados, em relação ao sentimento que expressam, em “positivo”, “negativo” ou “neutro”. A API apache OPENNLP para processamento de linguagem natural foi utilizada para separar tokens com suas respectivas classes gramaticais identificadas.

---

[1] Segundo o sítio “Meio e Mensagem”

[2] “Agevole” reponsável por auxiliar empresas quanto ao marketing online de seus produtos e serviços.

Nascimento et al. 2013, utiliza análise de sentimentos em tweets com foco em notícias, procurando descobrir o que as pessoas pensam e como se sentem em relação aos acontecimentos do dia a dia, onde analisa as opiniões para identificar se as pessoas tendem a classificar notícias como positiva ou negativa e, comenta que o Twitter se mostrou uma rica fonte de informação.

A apresentação de um modelo de reputação para identificar o nível emocional de indivíduos a partir de conteúdos postados no Facebook (SILVA, et al. 2016) caracteriza um indivíduo emocionalmente, como forma de estimar o quão confiável um indivíduo pode estar em problemas de tomada de decisão, onde identifica e compara informações compartilhadas por usuários no Facebook, com um conjunto de palavras afetivas, definidas pela ANEW-br.

Os textos podem conter mais que informação, podem expressar opinião e emoção. A análise de sentimento pode auxiliar na tomada de decisão, onde a classificação de um produto por exemplo, pode conter opinião pessoal, determinando se um produto pode ou não ser recomendado por uma pessoa, com base em opiniões de outras pessoas (DOSCIATTI, et al. 2012).

Neste trabalho são aplicados técnicas de mineração de dados em textos extraídos da rede social Twitter. Espera-se que, a identificação do estado emocional dos indivíduos que usam esta rede social possa ajudar na obtenção do grau de satisfação desses indivíduos quanto a assuntos relacionados à instituição, e com isso direcionar esforços/melhorias em questões que foram identificadas emocionalmente como tristeza, raiva ou decepção.

Para isso, é usado nesse trabalho a API do Twitter, uma biblioteca Tweepy 3.5 do Python para extração dos dados, códigos desenvolvidos em Python para o pré processamento dos textos, um modelo matemático proposto também citados por (PAIM, 2013) e (SILVA, et al., 2016); o método ANEW-Br será utilizado pois possui em sua base palavras na língua portuguesa e, as palavras anteriormente classificadas serão comparadas com as catalogadas pelo ANEW-Br.

Técnicas de mineração de textos são usadas para comparar o grau de similaridade das palavras postadas com o ANEW-Br, usando como base a metodologia usada em (SILVA, et al. 2014).

1.1 Justificativa

Com a crescente demanda pelo uso das redes sociais bem como o uso de textos não estruturados, a interpretação e o entendimento das avaliações dos usuários pelas empresas tem

se tornado uma tarefa difícil. Sabe-se que, afetos básicos de um indivíduo, como alegria, tristeza, raiva, etc, podem o levar a determinadas decisões. Por exemplo, um indivíduo no estado alegre pode expressar uma opinião positiva sobre algo ou alguém, já um indivíduo no estado triste pode expressar uma opinião contrária se comparado ao estado do indivíduo alegre.

É difícil definir o nível emocional de um indivíduo e o quanto seria confiável a informação gerada por ele para tomada de decisão levando em consideração seu estado emocional. A análise de sentimentos, tem um desafio referente à forma como a informação é interpretada, pois pode mudar o contexto e gerar descrédito principalmente em casos onde identifica-se sarcasmo e ironia que pode ser visto com mais detalhes no Capítulo 3. A habilidade de identificar sarcasmo e ironia vem recebendo a atenção de diversas pesquisas científicas na computação e em várias outras áreas do conhecimento em diálogos publicados na WEB, como por exemplo, na linguística (CHEANG and PELL, 2011) e na sociologia (BALL, 1965). No mundo real, as pessoas são capazes de compreender tais características em uma conversa devido a vários fatores contextuais tais como os gestos do locutor e seu tom de voz (GIBBS and COLSTON, 2007).

Nas redes sociais as pessoas demonstram o seu estado emocional com palavras, símbolos ou imagens; mostrando o quão felizes, tristes ou com raiva podem estar de algo ou alguém. A identificação do estado emocional de um indivíduo por meio de suas postagens em redes sociais, pode contribuir para fortalecer estratégias que venham a melhorar partes organizacionais com déficit, tanto de gestão, estrutural ou didádico pedagógica da instituição de ensino.

1.2 Objetivos

1.2.1 Objetivo Geral

Aplicar técnicas de mineração de dados em textos extraídos da rede social Twitter de uma Instituição de Ensino Técnico, Tecnológico e Superior.

1.2.2 Objetivos Específicos

- Levantamento bibliográfico;

- Definição das ferramentas para extração dos dados;
- Seleção e pré-processamento dos dados;
- Aplicação de uma técnica de mineração de texto;
- Interpretação e avaliação dos resultados;
- Redação da monografia.

1.3 Estrutura do Trabalho

Este trabalho está organizado da seguinte forma, como mostra o mapa mental na

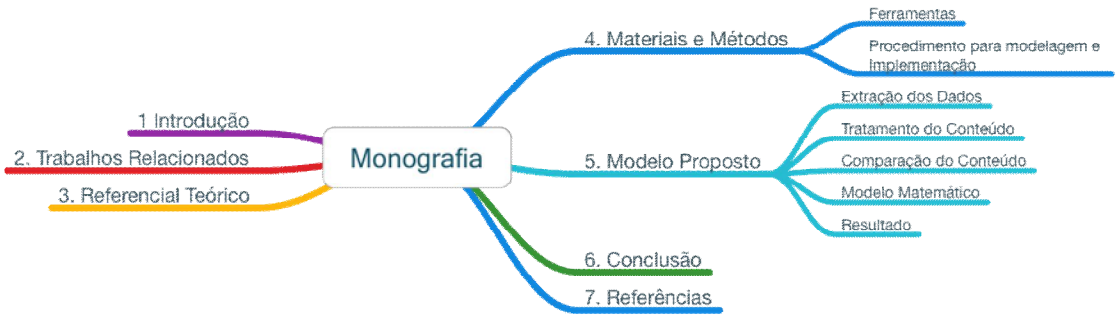


Figura 1, onde o capítulo 1 apresentou a introdução bem como a justificativa e os objetivos:

Figura 1: Mapa mental

O capítulo 2 apresenta alguns trabalhos relacionados ao tema; no capítulo 3 encontra-se o referencial teórico, onde apresenta os conceitos das técnicas que foram utilizadas no desenvolvimento dessa pesquisa; no capítulo 4
são descritos os materiais e métodos utilizados no projeto; o capítulo 5 é voltado para o modelo proposto, onde são demonstrados os meios para extração de dados, tratamento de conteúdo, comparação do conteúdo, modelo matemático e o resultado esperado e no capítulo 6 é apresentada a conclusão do trabalho.

2 TRABALHOS RELACIONADOS

Neste capítulo será descrito alguns trabalhos relacionados onde são direcionados à pesquisa, é citado a análise de sentimentos usando técnicas de mineração de textos. Algumas pesquisas possuem a mesma base de estudo do trabalho proposto, porém com o uso de técnicas e tecnologias diferentes.

Pesquisa / Artigo	Base de Dados	Materias/Métodos	Ano
PARIKH, Ravi; MOVASSATE, Matin. Sentiment analysis of user-generated twitter updates using various classification techniques. **CS224N Final Report**, p. 1-18, 2009.	Twitter	- Linguagem JAVA - API Twitter - Unigram Naive Bayes	2009
PAIM, Aldo. Mineração de Texto para a Análise de Sentimentos: Um estudo em Redes Sociais. Monografia Especialização Banco de Dados. UTFPR, Pato Branco.	Facebook	- API gráfica do Facebook - Delphi - JSON - ANEW-Br	2013
SILVA, William et al. IDENTIFICANDO EMOCÕES EM REDES SOCIAIS: Um estudo de caso no facebook. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 2, n. 10, p. 81-89, 2015.	Facebook	- Pentaho - API do Facebook - ANEW-Br - Linguagem C	2015
DOSCIATTI, Mariza Miola; FERREIRA, Lohann Paterno Coutinho; PARAISO, Emerson Cabrera. Identificando emoçoes em textos em português do brasil usando máquina de vetores de suporte em soluçao multiclasse. **ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil**, 2013.	Notícias extraídas de um jornal online	- Algoritmo SVM (Support Vector Machine) ou Máquina de vetores de suporte	2013
RIBEIRO, Lucas Braga. Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: estudo do impacto do pré-processamento. 2015.	Comentários sobre aplicativos móveis extraídos da Google Play	- GoogleMarketAPI - PLN - NLTK (stemmer)	2015

3 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados estudos sobre alguns métodos e tecnologias. O objetivo não é descrever ou implementar um projeto específico, mas analisar as potencialidades de novas tecnologias.

3.1 Mineração de Textos

A mineração de textos pode ser vista como uma extensão da área de Data Mining, focada na análise de textos (BARION, et al. 2015). Também chamada de Text Data Mining, Knowledge Discovery in Texts (KDT) (TAN, 1999) refere-se a extração de padrões ou conhecimentos interessantes e não triviais de documentos de texto. Já para (DORRE et al. 1999) mineração de texto se aplica a algumas funções analíticas de mineração de dados, mas também se aplica a funções analíticas de linguagem natural (LN) e técnicas de recuperação de informação - Information Retrival (IR).

As ferramentas de mineração de texto são utilizados para:

- Extrair informações relevantes de um documento - extrair as características (entidades) de um documento usando LN, IR e algoritmos métricas de associação (FELDMAN et al, 1998) ou correspondência padrão (DORRE et al. 1999);
- Encontrar tendência ou as relações entre pessoas / lugares / organizações, etc, através da agregação e comparar as informações extraídas dos documentos;
- Classificar e organizar documentos de acordo com o seu conteúdo (TKACH, 1998);
- Recuperação de documentos com base nos vários tipos de informações sobre o conteúdo do documento;
- Agrupar documentos de acordo com seu conteúdo (WAI-CHIU e FU, 2000).

Um sistema de mineração de texto é composto por 3 componentes principais (BEN-DOV e FELDMAN, 2005) :

1. **Information Fedders:** permite a ligação entre várias coleções textuais e os módulos de marcação. Este componente se conecta a qualquer site da web, streaming de origem (um novo feed), coleções de documentos internos e quaisquer outros tipos de coleções textuais;
2. **Intelligence Tagging:** componente responsável por ler o texto e separar (marcação) a informação relevante. Este componente pode realizar qualquer tipo de marcação sobre os documentos, tais como marcação de estatística (categorização e extração), semântica (extração de informações) e estrutural (extração do layout visual de documentos);
3. **Business Intelligence Suite:** um componente responsável por consolidar as informações de diferentes fontes, permitindo a análise simultânea de toda o quadro de informações.

A mineração de texto ganhou importância com o crescimento da internet e o aumento no volume de suas informações e dos mecanismos de busca (PINHEIRO, 2008).

Para (AGGARWAL et al., 2012) a pesquisa para a recuperação da informação refere-se em facilitar o acesso à informação, em vez de analisar as informações para descobrir padrões que é o principal objetivo da mineração de texto. O objetivo do acesso à informação é conectar as informações corretas com os usuários certos, no momento certo, com menos ênfase em processamento ou transformação de informações de texto.

Ainda segundo AGGARWAL, mineração de texto podem ser considerada como recurso ao acesso à informação onde auxilia os usuários a analisar as informações para a tomada de decisão. Há também muitas aplicações de mineração de texto onde o principal objetivo é analisar e descobrir quaisquer padrões, incluindo tendências e valores discrepantes em dados de texto. Tecnicamente, as técnicas de mineração focam os principais modelos, algoritmos e aplicações sobre o que se pode aprender a partir de diferentes tipos de dados de texto. Alguns exemplos de tais questões a seguir:

- ✓ Quais são os modelos supervisionados e não supervisionados para aprender a partir de dados de texto?
- ✓ Como são os problemas de agrupamento e de classificação tradicionais para dados de texto, em comparação com a literatura tradicional de banco de dados?
- ✓ Quais são as ferramentas úteis e técnicas utilizadas para a mineração dados de texto?

- ✓ O que são as técnicas matemáticas úteis que se deve saber, e que são repetidamente usadas no contexto de dados de texto em diferentes tipos de sistemas?
- ✓ Quais são os domínios de aplicação fundamentais em que são utilizadas tais técnicas de mineração, e como eles são efetivamente aplicados?

Técnicas foram estudadas e desenvolvidas com o objetivo de auxiliarem na extração de informações importantes, implícitas nas bases de dados (BARION, et. al. 2015).

As etapas pertencentes a descoberta do conhecimento em base de dados - Knowledge Discovery in Databases (KDD) são:

1. **Dados:** O KDD se baseia no armazenamento dos dados de forma estruturada;
2. **Seleção de Dados:** Após ter definido o domínio sobre o qual se pretende executar o processo de descoberta, a próxima etapa é selecionar e coletar o conjunto de dados ou variáveis necessárias;
3. **Processamento:** Esta etapa é também conhecida com pré processamento visando eliminar os dados que não se adequam às informações, com base nos algoritmos, ou seja, dados incompletos, problemas de definição de tipos, eliminação de tuplas repetidas, etc.;
4. **Transformação:** Nesta etapa os dados deverão ser armazenados adequadamente para facilitar na utilização das técnicas de mineração de dados;
5. **Mineração de Dados:** A atividade de descoberta do conhecimento é onde são processados os algoritmos de aprendizado de máquina e de reconhecimento de padrões. A maioria dos métodos de *Data Mining* são baseados em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, clusterização, modelos gráficos;
6. **Interpretação/Avaliação:** Nesta etapa final, os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas, porém devem ser apresentadas de forma que o usuário possa entender e interpretar os resultados obtidos.

As etapas da mineração de textos citadas por (ALVARES, et al. 2016) são descritas a seguir:

- **Extração da Informação:** a técnica de extração da informação tem como objetivo produzir uma saída estruturada a partir de textos (não estruturados). Essa técnica consiste da identificação de frase-chave e de relacionamentos no texto. E esta é uma

das dificuldades do uso da técnica: determinar, para cada problema ou domínio a ser analisado, as sequências relevantes.

- **Rastreamento de Tópicos (Topic Tracking):** É uma técnica que visa encontrar os documentos relacionados com determinados tópicos. Essa técnica tem grande potencial de exploração comercial e científico. A implementação da técnica consiste na identificação das palavra-chave do documento de comparação de relatividade da palavra-chave encontrada com os tópicos desejados. Existem inúmeras técnicas para a identificação de palavra-chave no texto. A maior parte delas se baseia na quantidade de aparições da palavra no corpo do texto. Algumas consideram ainda um sistema de pesos, atribuído as palavras, como forma de criar um ranking de persistência do documento.
- **Sumarização:** o objetivo é localizar palavras ou frases que tenham importância dentro de um texto ou conjunto de textos e, a partir daí, criar um resumo ou sumário do conteúdo. Esta técnica é especialmente útil para textos muito extensos. Desta forma o usuário determina através do sumário gerado, se o documento é ou não interessante para suas necessidades. O ponto principal é reduzir o tamanho do texto, mantendo as partes principais sem perder seu significado geral.
- **Categorização / Classificação:** determina a categoria ou classe a qual um documento pertence baseado em seu conteúdo. É a identificação dos temas principais de um documento localizando-o dentro de um conjunto pré-definido de tópicos. A implementação da técnica, é baseada em métodos de aprendizagem supervisionado. O uso principal dessa técnica é a indexação de documentos baseado em assunto. No entanto, a utilização da técnica é encontrada também em outros cenários na etapa de pré-processamento de textos.
- **Agrupamento (Clustering):** nesta técnica os documentos são agrupados de acordo com suas semelhanças e co-relacionamentos. Diferentemente do que o ocorre na categorização, no clustering não existe um conhecimento anterior das classes possíveis ou existentes. A descoberta dos grupos ocorre na execução do algoritmo, baseado unicamente no conteúdo dos documentos e referências auxiliares, como dicionários de domínio. A aplicação da técnica produz um tipo de conhecimento que precisa ser avaliada por um especialista do domínio. Vários algoritmos utilizados na versão da técnica aplicada a mineração de dados estruturados também se aplicam , com pequenas adaptações, a aplicação em textos não estruturados. O k-means é um

exemplo destes algoritmos. Na descoberta de conhecimento em textos, no lugar de atributos de uma tabela de banco de dados relacionais, utilizam-se estatísticas de aparições de termos ou frases nos documentos. Os relacionamentos entre termos lexicamente diferentes, porém semanticamente similares, também podem ser considerados para melhorar os resultados da técnica.

- **Acoplamento de conceitos (concept linkage):** refere-se a técnica de identificação de conexões entre documentos baseado na identificação dos conceitos compartilhados entre os mesmos. O objetivo desta técnica é auxiliar os usuários a encontrar informações que normalmente não conseguiram utilizando ferramentas de busca tradicionais. Um sistema de descoberta de conhecimento em texto comum poderia facilmente encontrar o relacionamento entre os tópicos X e Y e entre Y e Z. Um sistema que implemente a técnica de concept linkage, poderia apresentar diretamente o relacionamento entre os tópicos X e Z. Essa capacidade faz destes tipos de sistema especialmente úteis para algumas áreas de conhecimento, como biomedicina, onde o grande volume de produções científicas torna complexa a identificação desses relacionamentos de forma manual pelos pesquisadores. A pesquisa de um tratamento para uma determinada doença, poderia consumir várias horas de pesquisa e relacionamento entre documentos do mesmo tema. O suporte ferramental da técnica de concept linkage permite a identificação destes tipos de relacionamento de forma automática.
- **Visualização de informação:** consiste da apresentação de um grande conjunto de fontes textuais em uma hierarquia ou mapa. Além de apresentar os documentos, a implementação desta técnica oferecem resursos de navegação e busca. A ideia desta técnica é oferecer uma visualização esquemática dos assuntos, e o relacionamento entre estes, permitindo ao usuário uma forma visual de analisar o conteúdo de documentos. A implementação consiste da aplicação de três passos básicos: o primeiro é a preparação dos dados, pode consistir da limpeza dos dados, ou de processos de indexação e sumarização; o segundo passo é a análise e extração dos dados, o objetivo é criar a informação que será utilizada pelo último passo do processo, a construção da apresentação dos dados.

3.1.1 IRONIA x SARCASMO

Para melhor entendimento (GONÇALVES) cita a diferenciação entre sarcasmo e ironia brevemente com os conceitos desses dois temos:

• Sarcasmo: No sarcasmo, há um uso de instrumentos linguísticos indiretos para a ridicularização ou zombaria, muitas vezes considerado grosseiros e ofensivos, sendo utilizados para fins destrutivos (Singh, 2012). Um exemplo de sarcasmo pode ser visto no tweet “Super glad I got a sinus infection during finals week! #sarcasm”, em que é contrastado um sentimento de agradecimento com o surgimento de um contratempo negativo em um período possivelmente conturbado.

• Ironia: A ironia pode ser considerada como uma discordância, ou incoerência, entre o que se diz e o que se entende, ou do que se espera e do que realmente ocorre (Singh 2012). Esse tipo de mensagem geralmente vem acompanhado de um tom de brincadeira e possui menor peso ofensivo do que o sarcasmo. Um exemplo de ironia pode ser notado no tweet “Steve Jobs did not allow his kids to use iPads #irony”, onde, em um tom engraçado o usuário encontra uma possível contradição de uma situação.

3.2 Softwares para Text Minig

Atualmente existem vários softwares para mineração de textos, cada um com suas especificidades, vantagens e desvantagens. A seguir alguns softwares serão apresentados:

3.2.1 TextAnalyst

É uma ferramenta usada na descoberta de conhecimento em textos, processo conhecido como Mineração de texto (Text Mining), baseado na tecnologia de redes neurais. O software pode ser utilizado também no processo de descoberta de conhecimento de textos na WEB, sendo necessário salvar as páginas da web no formato de arquivo texto, pois a entrada para o TextAnalyst é um texto, do qual as informações relevantes são extraídas (DE MAGALHÃES, 2008).

As principais características da ferramenta TextAnalyst são:

- Para cada conceito são obtidos dois valores que representam os pesos semânticos: dos

conceitos em relação ao conceito “pai” e os pesos semânticos dos conceitos em relação ao documento.

- Permite, também, especificar determinadas palavras-chave, caso não tenha certeza se a ferramenta irá recuperá-las a partir do texto, possivelmente devido ao seu baixo peso semântico, as palavras-chave podem ser adicionadas pelo usuário com um dicionário externo.
- Permite criar sumários dos textos de entrada baseando-se em textos semânticos.
- Os resultados podem ser exportados para um arquivo no formato HTML ou para um formato compatível com a planilha do Excel.

3.2.**2 WEKA**

Weka é um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamado a partir de seu próprio código Java; contém ferramentas para pré-processamento de dados, classificação, regressão, clustering, regras de associação, e visualização. É também adequada para o desenvolvimento de novos sistemas de aprendizagem máquina. Weka é um software *open source* emitido sob a GNU General Public License (HALL et al.,2009).

WEKA surgiu fora da necessidade de ser capaz de aplicar o aprendizado de máquina para conjuntos de dados do mundo real de uma forma que promove uma "what if? ..." ou abordagem exploratória. Cada implementação do algoritmo de aprendizado de máquina requer que os dados estejam presentes em seu próprio formato, e tem sua própria maneira de especificar parâmetros e saída. O sistema WEKA foi concebido para proporcionar uma série de técnicas de aprendizagem de máquina ou sob regimes de uma interface comum, de modo que possam ser facilmente aplicados a esses dados num método consistente. Essa interface deve ser suficientemente flexível para incentivar a adição de novos esquemas, e simples o suficiente para que os usuários apenas se preocupem com a seleção de recursos na análise de dados e quais os meios de saída, ao invés de como usar um esquema de aprendizagem de máquina. (GARNER,1995).

3.3 Algoritmos para Mineração de Dados

Serão descritos alguns algoritmos e sua aplicação em mineração de dados. Os algoritmos citados apenas ilustram, mas não esgotam o universo de métodos de mineração de dados disponíveis. Cada método de mineração de dados requer diferentes necessidades de pré processamento (MORIK,2000). Tais necessidades variam em função do conjunto de dados em que o método será utilizado. Devido a vasta diversidade de métodos de pré-processamento de dados, são várias as alternativas possíveis de combinações entre métodos. A escolha dentre estas alternativas pode influenciar na qualidade do resultado do processo de KDD (MORIK, 2000; ENGELS, 1996; ENGELS et al., 1997; BERNSTEIN et al., 2002).

Na mineração de dados, são definidas as técnicas e os algoritmos a serem utilizados no problema em questão. Redes neurais (HAYKIN,1999), algoritmos genéticos (DAVIS,1990), modelos estatísticos e probabilísticos (MICHIE et al.,1994) são exemplos de técnicas que podem ser utilizadas na etapa de mineração de dados.

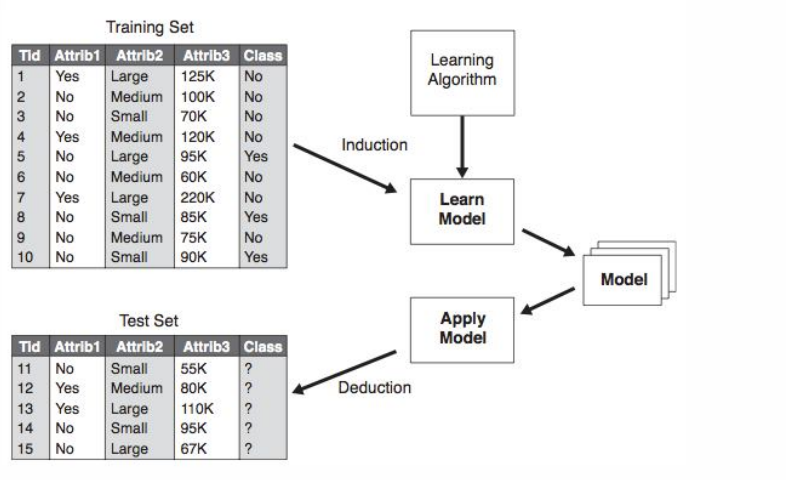
Considerando a evolução da análise de sentimentos, utilizando tokens ou informações extraídas de uma opinião, as pesquisas em mineração de opiniões para análise de sentimento podem ser agrupadas em quatro campos (CAMBRIA,2013):

- Palavras-chave e afinidade léxica: classifica o texto de acordo com a presença de palavras sem sentido ambíguo, tais como “feliz”, “triste” e “medo”. Além de detectar palavras óbvias, também atribui a outras palavras uma “afinidade” com um sentimento.
- Aprendizado de máquina: utiliza algoritmos de aprendizado de máquina, como Naive Bayes e SVM, para classificar um texto. Nesse caso, o sistema, além de aprender a importância de uma palavra-chave óbvia, considera outras palavras que podem ser fundamentais, além da pontuação e da frequência.
- Métodos estatísticos: esses métodos calculam a polaridade de uma palavra baseada na concorrência da mesma com palavras que possuem a mesma orientação.
- Baseado em conceitos: usam ontologias ou redes de palavras-chave para realizar a análise textual. Podem analisar expressões que não possuem uma emoção explícita, mas estão relacionadas a um sentimento implicitamente.

3.3.1 Resolução de um problema de Classificação

Técnicas de classificação é uma abordagem sistemática para construção de modelos de classificação a partir de um conjunto de dados de entrada. (TAN et al., 2009).

A Figura 2 mostra uma abordagem geral para resolver problemas de classificação, um conjunto de treinamento consistindo de registros cujos rótulos sejam conhecidos devem ser



fornecidos.

Figura 2: Abordagem geral para a construção de um modelo de classificação

A maioria dos algoritmos de classificação procura modelos que atinjam a maior precisão ou, equivalentemente, a menor taxa de erro quando aplicados ao conjunto de testes.

4.2.1 CRISP-DM

A metodologia CRISP-DM segundo (WIRTH et al, 2000), sugere um ciclo de fases que podem ser seguidos na resolução de problemas, de forma independente da indústria cujos dados são relacionados. A metodologia é apresentada na forma de um modelo hierárquico de processos, apresentando quatro níveis de abstração (do mais genérico para o mais específico): fases, tarefas genéricas, tarefas especializadas e instâncias de processos.

A Figura 3 apresenta a sequência e relação das fases da metodologia também citado por (RIBEIRO, 2015), que em tradução livre são: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e utilização. As fases e suas aplicações são explicadas a seguir.

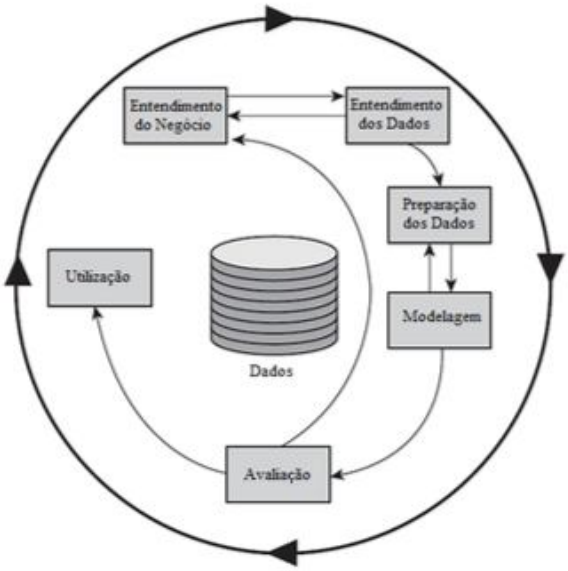


Figura 3: Descrição das fases do CRISP-DM

O entendimento do negócio (BusinessUnderstanding) se refere à etapa inicial da mineração de dados. Essa fase focaliza no entendimento do problema sob uma ótica negocial e então na conversão desse conhecimento em uma definição de um problema de mineração de dados. É elaborado um projeto direcionado em atingir determinados objetivos, também elencados nessa etapa, e são desenvolvidos cronogramas e listas de requisitos que poderão ser necessários durante toda a sequência do projeto. Durante esta etapa foram estudados os comentários dos seguidores da instituição de ensino no Twitter, a organização dos tweets e a comparação das palavras com as palavras catalogadas no ANEW-Br verificando a classe em que a mesma se encontrava.

3.4 Extração de Personalidade por meio de Textos

O trabalho de (LUNARDI, 2016, apud DAS e CHEN, 2001) foi um dos primeiros a analisar o sentimento das pessoas com dados da web, e utilizou o termo *extração de sentimento* para capturar a influência da opinião de indivíduos no domínio de finanças. Já (PANG; LEE; VAITHYANATHAN, 2002), utilizam o termo *classificação de sentimentos* para avaliar documentos considerando o sentimento geral de uma opinião, classificando-as

como positivas ou negativas. Outro trabalho inicial é o de Turney (TURNEY, 2002) que visa classificar opiniões como *recomendadas* ou *não recomendadas* (em inglês, *thumbs up* e *thumbs down*). Apenas em Nasukawa e Yi (NASUKAWA e YI, 2003) o termo *análise de sentimentos* é empregado, e assim como em (PANG; LEE; VAITHYANATHAN, 2002), os autores introduzem uma pesquisa para classificar uma opinião como positiva ou negativa.

3.4.1 Léxicos Afetivos

No trabalho de (PAIM, 2016) são reunidos léxicos afetivos que permitem utilizar não somente palavras que se referem diretamente a emoções, mas que implicam em diversas outras formas. Na tarefa de reconhecimento de personalidade, em que léxicos afetivos são usados, a abordagem baseada em dicionário é comumente utilizada, devido ao fato de ser livre de orientações específicas de contexto, existe ainda a abordagem baseada em corpus, onde o léxico é construído a partir de termos emocionais rotulados manualmente e novos termos emocionais são acrescentados, atualmente não há a disponibilidade de léxicos dessa abordagem em português. (PAIM, 2016) descreve ainda alguns léxicos afetivos (abordagem em dicionário) utilizados na literatura:

- **SentiStrength**. Combina uma abordagem baseada em léxico com regras linguísticas mais sofisticadas como erros ortográficos, pontuação e uso de emotions. Além de reportar a polaridade emocional (positivo e negativo) da palavra, pode-se ainda extrair a escala trinária (negativo, neutro e positivo)

- Affective Norms of English Words (ANEW). Pesquisadores do National Institute of Mental Health (NIMH), na Universidade da Flórida, desenvolveram diferentes conjuntos de estímulos emocionais e que fornecem classificações normativas de valência, alerta e dominância, são eles: International Affective Picture System (IAPS), International Affective Digital Sounds (IADS), Affective Norms for English Text (ANET) e Affective Norms for English Words (ANEW). Seu conjunto de palavras é composto por 1.034 vocábulos com medidas para três dimensões emocionais. A primeira dimensão diz respeito à valência, que consiste na avaliação do quão agradável ou desagradável é a percepção de uma palavra, isto é, a assimilação de um indivíduo a um vocábulo que provoca uma emoção discreta de felicidade pode ser classificada por uma valência agradável. A segunda dimensão é chamada de alerta, que avalia o quão estimulada ou relaxada é a percepção de uma palavra, como por exemplo, palavras que evocam raiva podem ser classificadas com alerta alto. A última dimensão,

chamada de dominância, consiste na análise do quão em controle de uma palavra ou dominado por ela. Considerando sua utilidade e relevância para o meio científico, o ANEW foi traduzido para uma versão em português, sofrendo adaptação e normalização para o idioma brasileiro.

A norma brasileira para o Affective Norms for English Words (ANEW-Br) (KRISTENSEN et al., 2011) teve como objetivo obter medidas de valência e alerta para um conjunto de 1.046 palavras em português, realizando um estudo de tradução, adaptação e normatização do ANEW. A base textual do Anew-Br possui palavras com valores de valência e alerta compreendidas no intervalo de 1 a 9.

Estudo correlatos psicofisiológicos indicam que julgamentos de alerta estão associados a variações na condutância elétrica da pele e julgamentos de valência estão associados a variações na contração de músculos faciais (KRISTENSEN et al., 2011).

3.5 Redes Sociais

A *Internet* facilitou as relações pessoais e o fluxo de informações entre redes de interesses comuns. Hoje é possível que amigos se encontrem no Facebook, no Twitter e em tantos outros “espaços virtuais” e, desse modo, comuniquem-se e estabeleçam uma rede de amizade, trabalho ou estudo. Interações estas denominadas na *Internet* de redes sociais, e hoje impulsionadas por diversos sites e ferramentas exclusivos para este fim (LUNARDON, 2013).

As pessoas estão inseridas na sociedade por meio das relações que desenvolvem durante toda a sua vida, desde o âmbito familiar até o trabalho; enfim, as relações que as pessoas desenvolvem e mantém é que fortalecem a esfera social. As redes sociais constituem uma das estratégias subjacentes utilizadas pela sociedade para o compartilhamento da informação e do conhecimento, mediante as relações entre atores que as integram (MARTELETO, 2001).

3.5.1 Twitter

Os microblogs são serviços mais recentes da WEB 2.0 voltados pra comentários curtos, rápidos e com atualizações constantes, que são acompanhados em forma de stream. Você decide quem deseja seguir e também filtrar quem o segue. O Twitter criado em março de

2006 é praticamente sinônimo de microblog, além de transmitir informações, seguir especilistas em determinados assuntos, compartilhar links e cobrir e acompanhar eventos (mesmo para quem não está lá presencialmente), o Twitter tem sido também utilizado criativamente para diferentes tipos de interações e discussões (MATAR, 2013).

3.5.2 Facebook

O The Facebook, criado por Mark Zuckerberg, estudante de Harvard, em janeiro de 2004 , era um serviço para conectar estudantes da universidade entre si e, em apenas 24 horas, mil pessoas se associaram a esse projeto e em um mês metade dos alunos já tinham criado um perfil. Em pouco tempo o serviço expandiu para outras universidades: Yale e Stanford. Em 2005, renomeado para Facebook, a ideia de conectar pessoas mediante o uso de um perfil, atualizações de estado desse perfil e utilização de foruns estava consolidada. Mais de 517 minhões de pessoas em 212 países diferentes juntaram-se a rede Facebook, num período de tempo surpreendentemente curto de seis anos. (ALVIM, 2011).

4 MATERIAS E MÉTODOS

Neste capítulo será descrito os métodos utilizados para chegar ao objetivo do projeto bem como as ferramentas utilizadas para o desenvolvimento do mesmo.

4.1 Ferramentas

As ferramentas utilizadas para o desenvolvimento, mineração de textos e análise de sentimentos foram as seguintes:

- Linguagem de programação Python
- API do Twitter
- Biblioteca Tweepy 3.5
- ANEW-Br

4.1.1 Linguagem de programação Python

A linguagem Python segundo (SUMMERFIELD, 2012) é bastante popular, por ser simples para ler e escrever; é uma linguagem multiplataforma e que pode ser usado para programar em paradigma procedural ou orientada a objetos. Possui uma completa biblioteca-padrão que permite inúmeras possibilidades com uma ou poucas linhas de código. Lançada por Guido van Rossum em 1991 como software livre, pode ser utilizada gratuitamente.

Segundo o site http://www.tiobe.com/tiobe-index//, a linguagem Python cresceu no ranking das linguagens mais utilizadas em 2016 ficando em quarto lugar.

A linguagem Python foi utilizada para desenvolver códigos para extrair dados do Twitter, remover caracteres especiais e espaços em branco do arquivo extraído, e desenvolver o algoritmo criado por (SILVA, 2014).

4.1.2 API do Twitter

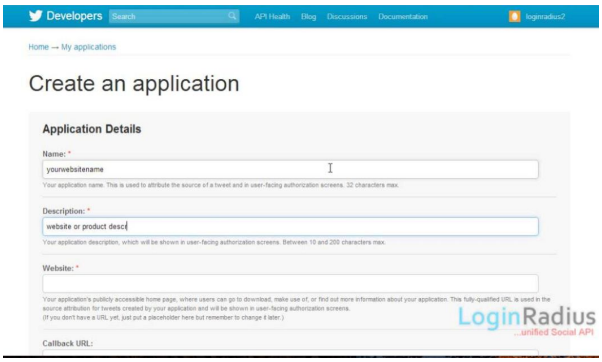
A rede social Twitter é a base de dados onde foram extraídos tweets públicos relacionados a instituição de ensino federal e, para isso foi necessário criar uma aplicação vinculada a conta do Twitter e gerar keys secrets (chaves secretas) que consiste em uma consumer key, consumer secret além dos access token e acces token secret, diante dessas informações geradas automaticamente, foi possível criar uma autenticação junto ao Twittter e desenvolver o código para extrair dados.

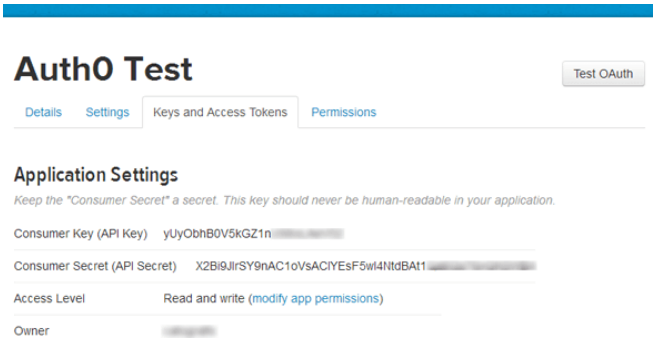
Figura 4: Imagem ilustrativa fonte: https://i.ytimg.com/vi/5PUC9yGS4RI/maxresdefault.jpg

Figura 5: Imagem ilustrativa fonte: https://cdn.auth0.com/docs/media/articles/connections/social/twitter/twitter- api-4.png

4.1.3 Tweepy 3.5

[3]Tweepy é uma biblioteca de acesso ao Twitter para o Python, acesse o site para fazer o download. O exemplo a seguir faz o download dos tweets da linha do tempo e mostra cada





um dos seus textos no console.

[3] www.tweepy.org

```
API.home_timeline([since_id] [,max_id] [,count] [,page])
```

Quadro 2: [4]Código Python

tweepy.api — Twitter API wrapper

```
class API([auth_handler=None] [,host='api.twitter.com'] [,search_host='search.twitter.com'] [,
cache=None] [,api_root='/1'] [,search_root=''] [,retry_count=0] [,retry_delay=0] [,
retry_errors=None] [,timeout=60] [,parser=ModelParser] [,compression=False] [,
wait_on_rate_limit=False] [,wait_on_rate_limit_notify=False] [,proxy=None])
```

```
api = tweepy.API(auth)

public_tweets = api.home_timeline()
for tweet in public_tweets:
    print tweet.text
```

O Twitter requer solicitação para usar o OAuth para autenticação, o mesmo foi apresentado anteriormente. A classe API fornece acesso a todos os métodos da API RESTful do Twitter. Cada método pode aceitar vários parâmetros e retornar respostas.

Figura 6: Twitter API

Algumas informações sobre esse metódos:

API do Twitter: esta classe fornece um wrapper para a API

Métodos da linha do tempo: a imagem a seguir retorna os 20 status mais recentes, incluindo retweets, postados pelo usuário de autenticação e amigos desse usuário. Este equivale a /timeline/home na WEB.

Figura 7: tweepy time_line

Statuses_lookup: retorna os objetos do tweet completos para até 100 tweets por solicitação, especificados pelo parâmetro ID.

Figura 8: tweepy statuses_lookup

[4] código disponível em www.tweepy.org

Acessando o sítio da biblioteca na *Internet*, é possível encontrar vários métodos que podem ser úteis a sua aplicação.

4.1.4 ANEW-Br

O ANEW-Br é uma base textual utilizada por esse estudo, contendo 1046 palavras que

```
API.statuses_lookup(id[, include_entities] [, trim_user] [, map])
```

foram traduzidas e adaptadas do Affective Norms for English Words (ANEW) para o português brasileiro.

O ANEW consiste em um conjunto de palavras com medidas para três dimensões emocionais. A primeira dimensão, chamada de valência, consiste na avaliação do quão agradável ou desagradável um estímulo é percebido. A segunda dimensão, chamada de alerta, consiste na avaliação do estimulado ou relaxado um estímulo nos deixa. A terceira dimensão, chamada dominância, consiste na avaliação do quão em controle de um estímulo ou dominado por ele nós nos percebemos.

O ANEW tem se mostrado uma ferramenta relevante na pesquisa sobre emoção, inspirando sua adaptação completa ou parcial em vários países.

Para as 1046 palavras em português, foi utilizada a escala Self-Assessment Manikin (SAM), para avaliação da emocionalidade do conjunto final.

A opção de utilizar o SAM como escala de avaliação emocional foi baseada em suas propriedades psicométricas. Esses resultados indicam que o SAM possui boas características psicométricas e apontam para a adequação da teoria dimensional da emoção.

4.2 PROCEDIMENTO PARA MODELAGEM E IMPLEMENTAÇÃO

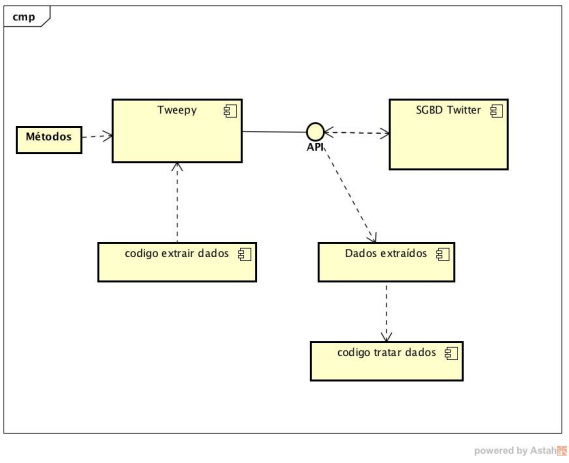
O trabalho consiste em uma tarefa de mineração de dados e, portanto foi utilizado como modelo de referência as fases do Cross-Industry Standard Process for DataMining - CRISP-DM, e o diagrama de componentes da UML citado a seguir para representar o projeto.

4.2.1 Diagrama de Componentes

O diagrama de componentes, descrito por (GUEDES, 2011) está associado à linguagem de programação que será utilizada para desenvolver o sistema modelado. Esse diagrama representa os componentes do sistema quando o mesmo for implementado em termos de módulos de código-fonte, bibliotecas, formulários, arquivos de ajuda, módulos executáveis, e determina como tais componentes estarão estruturados e irão interagir para que o sistema funcione de maneira adequada.

Seguindo o modelo proposto, a Figura 9 a seguir mostra o diagrama de componentes do projeto proposto:

Figura 9: Diagrama de Componentes



5 MODELO PROPOSTO

Este capítulo apresenta o modelo proposto, onde temos como objetivo identificar o nível emocional dos usuários do Twitter que mencionam a referida instituição de ensino em seus posts. Para tal, inicialmente foi realizado a extração dos dados na *Internet*, coletando os tweets públicos, posteriormente, inicia-se a fase do tratamento dos dados coletados e na sequência, a comparação dos dados com as palavras catalogadas no ANEW-Br. Com as palavras identificadas junto ao catálogo da ANEW-Br, aplica-se o modelo matemático desenvolvido e citado por (PAIM, 2013) e inserido nesse projeto.

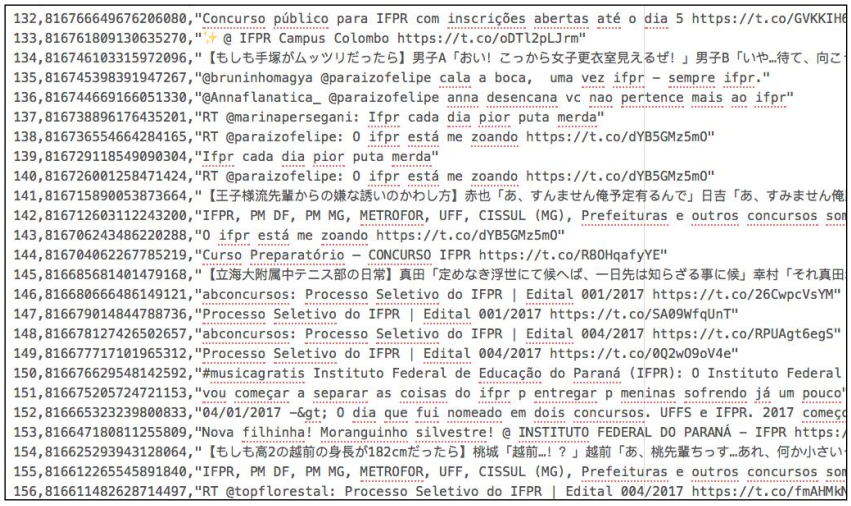
5.1 Extração de dados

Para a extração dos dados foi utilizado a biblioteca Tweepy 3.5 do Python, os dados extraídos foram salvos em um arquivo com extensão .csv com atributos como tweet_id e tweet_text, não sendo necessárias para a realização da mineração de textos e análise de

```
import tweepy
import re
import csv

consumer_key = "xxx"
consumer_secret = "xxx"
access_token = "xxx"
access_token_secret = "xxx"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

criterio = "busca"

tweets = api.search(criterio)

resultado = open('arquivo.csv','w')

i = 1

while tweets:
    print("\nEncontrei %d tweets\n" % (len(tweets)))

    for tweet in tweets:
        print("%d,%d,\"%s\"\n" % (i,tweet.id,tweet.text))
        i = i + 1
        last_id = tweet.id
        resultado.write("%d,%d,\"%s\"\n" % (i,tweet.id,tweet.text))

    tweets = api.search(criterio, max_id=last_id-1)

print("successful!!!")
```

Quadro 3: Código Python: Extrair dados do Twitter

sentimentos informações como o username, date e etc.



```
132,816766649676206080,"Concurso público para IFPR com inscrições abertas até o dia 5 https://t.co/GVKKIH6
133,816761809130635270,"✨ @ IFPR Campus Colombo https://t.co/oDTl2pLJrm"
134,816746103315972096,"【もしも手塚がムッツリだったら】男子A「おい! こっから女子更衣室見えるぜ! 」男子B「いや…待て、向こ
135,816745398391947267,"@bruninhomagya @paraizofelipe cala a boca,  uma vez ifpr - sempre ifpr."
136,816744669166051330,"@Annaflanatica_ @paraizofelipe anna desencana vc nao pertence mais ao ifpr"
137,816738896176435201,"RT @marinapersegani: Ifpr cada dia pior puta merda"
138,816736554664284165,"RT @paraizofelipe: O ifpr está me zoando https://t.co/dYB5GMz5mO"
139,816729118549090304,"Ifpr cada dia pior puta merda"
140,816726001258471424,"RT @paraizofelipe: O ifpr está me zoando https://t.co/dYB5GMz5mO"
141,816715890053873664,"【王子様流先輩からの嫌な誘いのかわし方】赤也「あ、すんません俺予定有るんで」日吉「あ、すみません俺
142,816712603112243200,"IFPR, PM DF, PM MG, METROFOR, UFF, CISSUL (MG), Prefeituras e outros concursos som
143,816706243486220288,"O ifpr está me zoando https://t.co/dYB5GMz5mO"
144,816704062267785219,"Curso Preparatório - CONCURSO IFPR https://t.co/R8OHqafyYE"
145,816685681401479168,"【立海大附属中テニス部の日常】真田「定めなき浮世にて候へば、一日先は知らざる事に候」幸村「それ真田
146,816680666486149121,"abconcursos: Processo Seletivo do IFPR | Edital 001/2017 https://t.co/26CwpcVsYM"
147,816679014844788736,"Processo Seletivo do IFPR | Edital 001/2017 https://t.co/SA09WfqUnT"
148,816678127426502657,"abconcursos: Processo Seletivo do IFPR | Edital 004/2017 https://t.co/RPUAgt6egS"
149,816677717101965312,"Processo Seletivo do IFPR | Edital 004/2017 https://t.co/0Q2wO9oV4e"
150,816676629548142592,"#musicagratis Instituto Federal de Educação do Paraná (IFPR): O Instituto Federal
151,816675205724721153,"vou começar a separar as coisas do ifpr p entregar p meninas sofrendo já um pouco"
152,816665323239800833,"04/01/2017 -&gt; O dia que fui nomeado em dois concursos. UFFS e IFPR. 2017 começo
153,816647180811255809,"Nova filhinha! Moranguinho silvestre! @ INSTITUTO FEDERAL DO PARANÁ - IFPR https:/
154,816625293943128064,"【もしも高2の越前の身長が182cmだったら】桃城「越前…！？」越前「あ、桃先輩ちっす…あれ、何か小さい
155,816612265545891840,"IFPR, PM DF, PM MG, METROFOR, UFF, CISSUL (MG), Prefeituras e outros concursos som
156,816611482628714497,"RT @topflorestal: Processo Seletivo do IFPR | Edital 004/2017 https://t.co/fmAHMkN
```

Figura 10: Trecho dos dados extraídos

5.2 Tratamento do Conteúdo

```
SE
FOR
FALAR
IFPR
NÃO
AMOR
LINDA
IDÉIAS
GOSTO
INFÂNCIA
COXA
SACANEIA
AMIGAS
FESTAS
CABELOS
VENTO
GOSTA
SAFADÃO
HOJE
PÉSSIMO
DIA
BLOQUEAR
SIGAA
FEDERAL
VOCÊS
RECUSARIAM
ALUNA
```

```
def checkfilecontent():
    import re
    import csv

    myworldlist = []
    result = open('arquivo                )
    resultado = open('arqu                csv','w')
    for items in result.re                t():
        char = ""
        if char in "":
            items = items.                (None, '.%,:\"|/@!?<>}{#][')
            items = items.                '',"")
            items = items.                (None, '0123456789')
            items = items.
            items = items.upper()
            if items != '':
                myworldlist.append(items)
                print items

        resultado.write("%s\n" % items)

    print("Total no of lines read in is " + str(len(myworldlist)))
    return myworldlist
checkfilecontent()
```

Quadro 4: Código Python - Remover caracteres e espaços

A etapa de pré-processamento é responsável por selecionar os atributos relevantes para o processamento dos textos e remover as palavras irrelevantes (stopwords[5]), utilizando algoritmo adaptado na linguagem Python foi possível separar as palavras e remover caracteres especiais, espaços, preposições, pronomes etc.

Figura 11: Dados Processados

5.3 Comparação do Conteúdo

Depois de realizar o tratamento do conteúdo extraído do Twitter, obteve-se um catálogo de palavras para que as mesmas fossem comparadas junto ao catalogo do ANEW-Br, a fim de obter o estado emocional.

O catálogo de palavras não sofre nenhuma alteração nesse processo, tem como função

[5]São palavras que podem ser consideradas irrelevantes para o conjunto do resultados a ser exibido em uma busca realizada em uma search engine. Ex: as, e, os, de, para, com, sem, foi.

restrita em buscar palavras idênticas no conjunto do ANEW-Br.

A tarefa de comparação foi utilizado o algoritmo de (SILVA, 2014) executado na linguagem C e citado também por (PAIM, 2013) que utilizou Object Pascal (Delphi), sendo dele a atribuição de comparação e decisão do descarte das palavras ou não. O descarte acontece caso seja identificado que o vocábulo não é idêntico ao do ANEW-Br, assim a palavra em questão não sofre interferência na criação do modelo matemático, pois não há medida emocional na mesma.

Figura 12: Pseudocódigo, algoritmo para modelo de reputação proposto (SILVA, 2014)

```
ENTRADAS:
 ESTRUTURA
  Palavra: STRING;
  Valencia, Alerta: REAIS;
 FIM DA ESTRUTURA;
SomaValencia, SomaAlerta, Mv, MA: REAIS;
Vn: REAL; //Valor de valência de uma palavra n
An: REAL; //Valor de alerta de uma palavra n
qn: INTEIRO; //Quantidade de palavras em comum entre o Anew-br e o
arquivo do facebook
Arq_Facebook: ARQUIVO;//Arquivo com palavras de um individuo obtidos  do facebook
Arq_Anew: ESTRUTURA; //Base textual do Anew-br
```

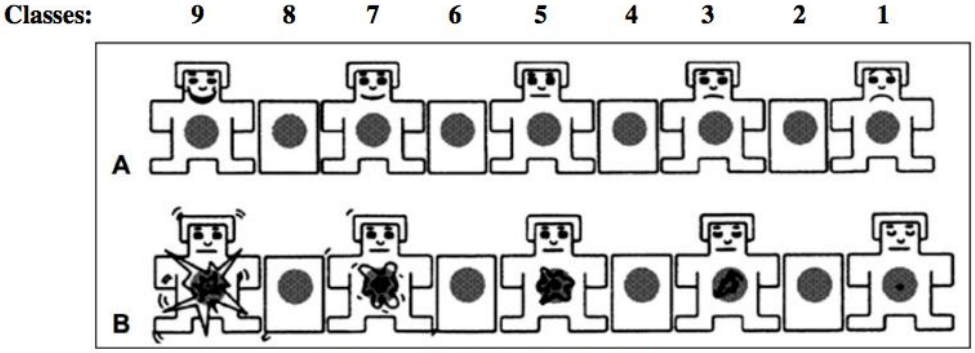
```
01   INICIO:
02   SomaValencia <- 0; SomaAlerta <- 0; qn <- 0;

03   Transfira os dados de Arq_Facebook e Arq_Anew para a RAM;
04   Leia a primeira palavra de Arq_Facebook;
05    ENQUANTO (Não EOF de Arq_Facebook)FAÇA//Enquanto não for final
      de arquivo
06      SE (token) ENTÃO
07        Elimine a primeira e última palavra;
08        Selecione a próxima palavra lida e consulte essa
          palavra em Arq_Anew;
09          SE (Encontrou) ENTÃO
10            SomaValencia <- SomaValencia + Vn;
11            SomaAlerta <- SomaAlerta + An;
12            qn <- qn + 1;
13            Leia a próxima palavra de Arq_Facebook;
14          FIM SE
15          SENÃO
16            Leia a próxima palavra de Arq_Facebook;
17          FIM SENÃO
18      FIM SE
19      SENÃO
20        Selecione a palavra e consulte a mesma em Arq_Anew;
21        SE (Encontrou) ENTÃO
22           SomaValencia <- SomaValencia + Vn;
23           SomaAlerta <- SomaAlerta + An;
24           qn <- qn + 1;
25           Leia a próxima palavra de Arq_Facebook;
26        FIM SE
27        SENÃO
28           Leia a próxima palavra de Arq_Facebook;
29        FIM SENÃO
30      FIM SENÃO
31    FIM ENQUANTO
32    SE (qn <> 0) ENTÃO
33      Mv <- SomaValencia/qn; //Conforme a equação 1
34      MA <- SomaAlerta/qn; //Conforme a equação 2
35    FIM SE
36    SENAO
37      Escreva: "Impossivel calcular as medias de valência e Alerta";
38    FIM SENÃO
```

5.4 Modelo Matemático

Para atingir o objetivo na construção de um modelo que tivesse a condição de medir o nível emocional de um indivíduo, foi utilizado o modelo matemático desenvolvido por (PAIM, 2013) que estima a média ponderada de valência e alerta.

A base textual do ANEW-Br possui palavras com valores de valência e alerta compreendidas no intervalo de 1 a 9. Sendo assim palavras que possuem valores baixos, próximos de 1, por exemplo, apresentam valência e alerta baixos, ou seja, desagradável e relaxado respectivamente. Já palavras com valores próximos de 9, apresentam valência e alerta altos, ou seja, agradável e estimulados respectivamente.

Sendo assim, segundo (PAIM, 2013) as palavras do ANEW-Br foram agrupadas em classes de 1 até 9 da seguinte forma: palavras com valores de valência compreendidas entre 1 e 1.99 pertencem a classe 1, entre 2 e 2.99 pertencem a classe 2, entre 3 e 3.99 classe 3 e



Fonte: Adaptado de Kristensen, 2011

assim sucessivamente. O mesmo foi feito para o alerta.

Figura 13: Classes de Valência e Alerta

Na Figura 13 as classes 1, 3, 5, 7, e 9 possuem a imagem de indivíduo que tenta representar as características físicas-sentimentais dessas classes, pois estudos correlatos psicofisiológicos indicam que a) julgamentos de alerta estão associados a variações na condutância elétrica da pele, ao passo que b) julgamentos de valência estão associados a variações na contração de músculos faciais, medida com eletromiografia facial (BRADLEY, M. M, LANG, P. J, 1999). As demais classes são consideradas classes intermediárias.

A partir disso, foram desenvolvidos modelos matemáticos que emergem um modelo de reputação capaz de estimar a média ponderada de valência (MV) e alerta (MA), conforme equações 1 e 2:

$$M_V = \frac{q_1 \times V_1 + \cdots + q_n \times V_n}{q_1 + \cdots + q_n} \quad (1)$$

$$M_A = \frac{q_1 \times A_1 + \cdots + q_n \times A_n}{q_1 + \cdots + q_n} \quad (2)$$

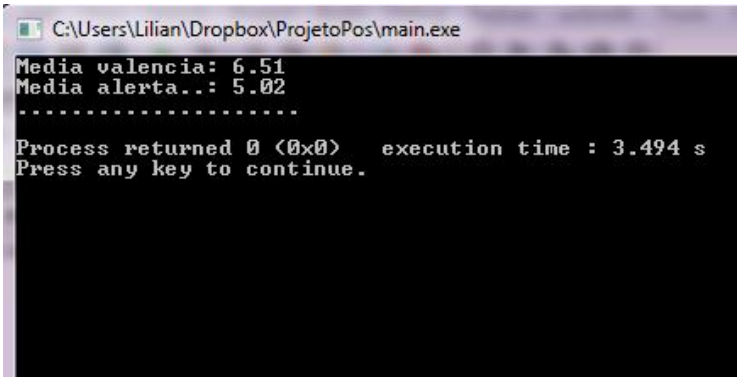
MV é a média ponderada para a valência; MA a média ponderada para o alerta; qi, para i = 1,…,n, a quantidade de vezes que uma palavra é encontrada; Vi, para i=1,…n, o valor de valência de uma palavra; e Ai, para i=1,…n, o respectivo valor de alerta de uma palavra.

5.5 Resultado

A pesquisa foi desenvolvida com base em posts de usuários do Twitter onde mencionavam a instituição de ensino no mês de dezembro, totalizando cem tuítes, somente informações declaradas como posts públicos foram extraídas.

O Twitter como outras redes sociais também possui regras de privacidade, para que fosse possível a extração dos dados mesmo sendo dados públicos, o microblog exige que para tal ação o usuário tenha uma conta no Twitter, e com essa conta, o mesmo crie uma aplicação onde terá permissão e autenticação necessária para extrair ou enviar posts usando API’s que o próprio Twitter disponibiliza.

Constatou-se com a aplicação, que os tweets que mencionam a instituição de ensino enquadram-se na classe de valência 6 e alerta 5, onde as médias ficaram com 6,51 e 5,02 respectivamente. Onde podemos afirmar que os posts não caracterizam desagrado emocional por parte dos usuários, e também não atingem nível emocional de extrema satisfação.



O modelo utilizado provou ser eficaz, capaz de medir o estado emocional de um indivíduo ou grupo. As etapas da mineração de textos foram empregadas, possibilitando adicionar novas funcionalidades em trabalhos futuros.

Figura 14: Resultado

6 CONCLUSÃO

Foi realizado leitura de vários trabalhos referentes a análise de sentimentos que abordam diversas formas de chegar ao resultado esperado. Todos os trabalhos que foram referência ao trabalho proposto usavam mineração de textos. A presente proposta, além de ater-se e atender aos conceitos de mineração de dados, buscou seu resultado com ênfase no ANEW, especificamente no ANEW-Br.

O Twitter é uma base de dados fortemente estudada na aplicação de mineração de dados, como análise de sentimentos e/ou opinião. Porém, a busca escolhida nessa base, a "instituição de ensino", limitou de certa forma o número de posts relevantes relacionados a ela, como exemplo o microblog limita seus posts a 140 caracteres, e muitas mensagens estavam em forma de emoticons ou com várias abreviações, o que gerou descarte de mensagens por parte do software.

O projeto utilizou modelos matemáticos e algoritmos já testados em outras pesquisas, buscando apresentar resultado satisfatório como já comprovado por outros estudos, o diferencial foi quanto a utilização da linguagem Python e bibiotecas específicas para conexão com Twitter e extração dos dados.

Para trabalhos futuros, esforços poderiam ser direcionados para o reconhecimento das palavras não catalogadas no ANEW-Br, onde essas tivesse um meio de serem reconhecidas em uma classe, a qual se identifica emocionalmente e, com isso gerar a média de valência e de alerta.

8 REFERÊNCIAS

<http://www.cs.waikato.ac.nz/ml/weka/index.html> Acesso em 27 out 2016.

AGEVOLE: **Estatisticas das redes sociais:** números que vão fazer você abrir os olhos <http://www.agevole.com.br/blog/redes-sociais/estatisticas-das-redes-sociais-numeros-que-vao-fazer-voce-abrir-os-olhos/>. Acesso em 02 out. 2016.

AGGARWAL, Charu C.; ZHAI, ChengXiang. **Mining text data**. Springer Science & Business Media, 2012.

ALVES, F.J. **Intodução à linguagem de programação Python.** Rio de Janeiro: Editora Ciência Moderna Ltda, 2013.

ALVIM, Luísa. **Impossível não estar no Facebook! O nascimento das bibliotecas portuguesas na rede social. CadERnOs BAd**, n. 1/2, 2011. <http://www.bad.pt/publicacoes/index.php/cadernos/article/view/737/736>Acesso em 16 out. 2016

BALL, D. W. (1965). **Sarcasm as sociation: The rhetoric of interaction**. Pages 190–198.

BARION, Eliana Cristina Nogueira; LAGO, Decio. Mineração de textos. **Revista de Ciências Exatas e Tecnologia**, v. 3, n. 3, p. 123-140, 2015.

BEN-DOV, Moty; FELDMAN, Ronen. Text mining and information extraction. In: **Data Mining and Knowledge Discovery Handbook**. Springer US, 2005. p. 801-831.

BERNSTEIN, Abraham; PROVOST, Foster; HILL, Shawndra. Intelligent Assistance for the Data Mining Process:.. In: **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**. 2002.

CAMBRIA, Erik et al. New avenues in opinion mining and sentiment analysis. **IEEE Intelligent Systems**, v. 28, n. 2, p. 15-21, 2013.

CARLANTONIO,L.M. **Novas Metodologias para Clusterização de Dados**. Dissertação de Mestrado, Engenharia Civil, Coppe, Universidade Federal do Rio de Janeiro,2001.

CHEANG, H. S. and PELL, M. D. (2011). **Recognizing sarcasm without language: A cross-linguistic study of english and cantonese**. page 19.

CORTES,C.,VAPNIK,V. **Support-Vector Networks, Machine Learning**,v.20. Kluwer Academic Publishers, 1995, p. 273-97.PINHEIRO, C.A.R..**Inteligência Analítica: Mineração de Dados e Descoberta Analítica**.Rio de Janeiro: Editora Ciência Moderna Ltda,2008.

CRUVINEL, Gustavo Warzocha Fernande. **Análise de Sentimentos em Redes Sociais Digitais: Uma aplicação no contexto político**. Disponível em: <http://www.academia.edu/7882935/Analise_de_Sentimentos_em_Redes_Sociais_Digitais_Uma_aplicacao_no_contexto_politico>. Acesso em: 07 de Outubro de 2014.

DAS, S. R; CHEN, M. Y. **Yahoo! For Amazon: Opinion Extraction from Small Talk on the Web**. Proceedings of the 8th Asia Pacific Finance Association Annual Conference, v. XXXIII, n. 2, p. 81–87, 2001.

DAVIS,L. **Handbook of Genetic Algorithms**, VNR Comp. Library, 1990.

DE MAGALHÃES, Lúcia Helena. **Uma análise de ferramentas para mineração de conteúdo de páginas Web**. 2008. Tese de Doutorado. UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. <http://wwwp.coc.ufrj.br/teses/mestrado/Novas_2008/teses/MAGALHAES_LH_08_t_M_int.pdf> Acesso em 14 out. 2016

DORIA, Marcela. **Redes Sociais são fortes candidatas a medalha de ouro.** <http://www.meioemensagem.com.br/home/opiniao/2016/05/25/rede-sociais-sao-fortes-candidatas-a-medalha-de-ouro.html >. Acesso em 02 out. 2016.

DORRE, J., GERSTL, P., and SEIFFERT, R. 1999. **Text mining: finding nuggets in mountains of textual data.** In Proceedings of KDD-99, 5th ACM International Conference on Knowledge Discovery and Data Mining (San Diego, US, 1999), pp. 398–401.

DOSCIATTI, M.D.; MARTINAZZO,B.; PARAISO, E.C.. **Identifying Emotions in Short Texts** for Brazilian portuguese. In. IV International Worshop on Web and Text Intelligence, 2012, Curitiba, Brazil. October/2012.p. 1-10

ENGELS, Robert; LINDNER, Guido; STUDER, Rudi. A guided tour through the data mining jungle. In: **KDD**. 1997. p. 163-166.

ENGELS, Robert. Planning tasks for Knowledge Discovery in Databases; Performing Task-Oriented User-Guidance. In: **KDD**. 1996. p. 170-175.

FELDMAN, R., FRESKO, M., Kinar, Y, Lindell, Y, Liphstar, O., Rajman, M., Schler, Y, and Zamir, O. (1998). **Text Mining at the Term Level.** Paper presented at the In Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, Nantes, France.

GARNER, Stephen R. et al. **Weka: The waikato environment for knowledge analysis.** In: Proceedings of the New Zealand computer science research students conference. 1995. p. 57-64.

GIBBS, R. W. and COLSTON, H. L. (2007). **Irony in language and thought: A cognitive science reader.** Psychology Press.

GONÇALVES, Pollyanna et al. **Bazinga! Caracterizando e Detectando Sarcasmo e Ironia no Twitter.**

GUEDES, Gilleanes T.A. **UML 2: Uma abordagem prática.** 2º ed. São Paulo: Novatec Editora, 2011.

HAYKIN, S. (1999). **Neural networks, a Comprehensive Foundation**. 2nd ed. Prentice Hall, Englewood Cliffs, New Jersey

KRISTENSEN, Christian Haag et al. Brazilian norms for the affective norms for English words. **Trends in psychiatry and psychotherapy**, v. 33, n. 3, p. 135-146, 2011.

LUNARDI,Alexandre de C. **Classificação Multiclasse De Textos Baseada Em Divisões Binárias Adaptadas Ao Domínio**.Dissertação de Mestrado.Universidade Federal Fluminense. <http://www2.ic.uff.br/PosGraduacao/Dissertacoes/722.pdf> Acesso em 28 out. 2016.

LUNARDON, Eliane Aparecida Dias, Sérgio Junqueira, and Pontificia Universidade Catolica do Paraná. Programa de Pós-graduação em Teologia. **As Redes Sociais Como Recurso Da Educação a Distância Na Formaçao Do Professor De Ensino Religioso** / Eliane Aparecida Dias Lunardon ; Orientador, Sérgio Rogério Azevedo Junqueira. 2013.

Mark HALL, Eibe FRANK, Geoffrey HOLMES, Bernhard PFAHRINGER, Peter REUTEMANN, Ian H. WITTEN (2009); **The WEKA Data Mining Software: An Update; SIGKDD Explorations**, Volume 11, Issue 1.

MARTELETO, Regina Maria. **Análise de redes sociais: aplicação nos estudos de transferência da informação**. Ciência da informação, v. 30, n. 1, p. 71-81, 2001.

MATTAR, João. **WEB 2.0 e Redes Sociais na Educação**. São Paulo: Artesanato Educacional,2013.

MICHIE, Donald; SPIEGELHALTER, David J.; TAYLOR, Charles C. Machine learning, neural and statistical classification. 1994.TAN,P.; STEINBACH,M.; KUMAR,V..**Introdução ao Data Mining: Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.

MORIK, Katharina. The representation race—preprocessing for handling time phenomena. In: **European Conference on Machine Learning**. Springer Berlin Heidelberg, 2000. p. 4-19.

NASCIMENTO, Paula; OSIEK, Bruno; XEXÉO, Geraldo. Análise de sentimento de tweets com foco em notícias. **Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.5329/RESI**, v. 14, n. 2, 2015.

NASUKAWA, T.; YI, J. **Sentiment Analysis : Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions**. 2nd International Conference on Knowledge Capture, p. 70–77, 2003.

PANG, B.; LEE, L.; VAITHYANATHAN, S. **Thumbs Up? Sentiment Classification Using Machine Learning Techniques**. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, n. July, p. 79–86, 2002.

RIBEIRO, Lucas Braga. **Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: estudo do impacto do pré-processamento**. 2015.

SILVA,W.; RIBEIRO,R.;TEIXEIRA,M.;ENENBRECK,F.. **Identificando Emoções em Redes Sociais**: um estudo de caso no facebook. Revista Eletrônica Científica Inovação e Tecnologia,Curitiba, volume 2, n. 10, 2014. Disponível em: <https://periodicos.utfpr.edu.br/recit/article/view/4308/Williana> Acesso em: 02 out. 2016.

SINGH, Raj Kıshor. **Humour, irony and satire in literature. Internatıonal Journal of English and Literature** (IJEL), v. 3, n. 4, p. 65-72, 2012.

SUMMERFIELD, Mark. **Programação em Python 3: Uma Introdução completa à Linguagem Python**. Rio de Janeiro: Alta Books, 2012.

TAN, Ah-Hwee et al. Text mining: The state of the art and the challenges. In: **Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases**. 1999. p. 65-70. <http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf> Acesso em: 11 out. 2016.

TKACH, D. (1998). "**Turning information into knowledge.**" a white paper from IBM.

TURNEY, P. D. **Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews**. Proceedings of the 40th annual meeting on association for computational linguistics, 2002.

WASSERMAN,Stanley; FAUST, Katherine. **Social Network Analysis:** Theory and Applications. Jan.2011.Disponível em: <http://www.asecib.ase.ro/mps/SocNet_TheoryApp.pdf>. Acesso em 18 set. 2016.

WILLIAMS, Kate; DURRANCE, Joan C. Social networks and social capital: Rethinking theory in community informatics. **The Journal of Community Informatics**, v. 4, n. 3, 2008.

WIRTH, Rüdiger, HIPP, Jochen. **Crisp-dm: Towards a standard process model for data mining**. 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000.