UNIVERSIDADE DO SAGRADO CORAÇÃO

DIEGO REGYS LOPES

SISTEMA PARA MINERAÇÃO DE SENTIMENTOS EM REDES SOCIAIS

BAURU

2014

DIEGO REGYS LOPES

SISTEMA PARA MINERAÇÃO DE SENTIMENTOS EM REDES SOCIAIS

Trabalho de conclusão de curso apresentado ao Centro de Ciências Exatas e Sociais Aplicadas como parte dos requisitos para obtenção do título de bacharel em Ciência da Computação, sob a orientação do Prof. Me. Patrick Pedreira Silva.

BAURU

2014

DIEGO REGYS LOPES

SISTEMA PARA MINERAÇÃO DE SENTIMENTOS EM REDES SOCIAIS

Trabalho de conclusão de curso apresentado ao Centro de Ciências Exatas e Sociais Aplicadas como parte dos requisitos para obtenção do título de bacharel em Ciência da Computação, sob a orientação do Prof. Me. Patrick Pedreira Silva.

Banca Examinadora:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Me. Patrick Pedreira Silva

Universidade do Sagrado Coração

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Esp. André Luiz Ferraz Castro

Universidade do Sagrado Coração

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Esp. Henrique Pachioni Martins

Universidade do Sagrado Coração

Bauru, \_\_\_\_\_\_de \_\_\_\_\_\_\_\_\_\_\_ de 2014.

Dedico primeiramente este trabalho e todo o percurso pessoal de aprendizado ao meu querido avô materno Manoel Lopo, um ser humano excepcional de extrema sabedoria e inteligência que foi e será meu norte por toda a vida. E depois aos meus pais e amigos pelas alegrias, tristezas e dores compartilhadas.

AGRADECIMENTOS

A esta universidade, seu corpo docente, direção e administração que proporcionaram um ambiente criativo, amigável e fértil para florescer não apenas o conhecimento racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por ter me feito aprender. A palavra mestre, nunca fará justiça aos professores dedicados aos quais sem nominar terão os meus eternos agradecimentos.

Aos meus pais, pelo amor, incentivo е apoio incondicional.

Aos meus amigos Camilla Penteado Farid, John Heverton Bezerra, e Rafaela Venerando, companheiros de trabalhos que fizeram parte da minha formação е que vão continuar presentes em minha vida com certeza.

Meus agradecimentos especiais ao amigo Guilherme Fainer Dalbem, que sem sua paciência e ajuda nos dias mais nebulosos este trabalho não estaria pronto.

"Saber muito não lhe torna inteligente. A inteligência se traduz na forma que você recolhe, julga, maneja e, sobretudo, onde e como aplica esta informação."

(Carl Sagan)

“O conhecimento passou a ser o principal fator de produção e geração de riquezas”

(Bill Gates)

**RESUMO**

Com o enorme volume de informações textuais geradas dia-a-dia torna-se crescente a necessidade de métodos e técnicas de extração e análise de conhecimento em textos. Hoje em dia quase não há uma transação que não seja feita por computadores e que gere registros que incluem informações valiosas como tendências e padrões, que analisados da forma correta, auxiliam na tomada de decisão. Considerando este contexto, o objetivo desta pesquisa foi investigar métodos de análise automática de sentimentos, propondo o desenvolvimento de um software. A ferramenta foi implementada tendo por base um corpus de comentários de usuários selecionados, a priori, manualmente de duas redes sociais (Facebook e Twitter), contendo #’s(hastags) vinculadas com palavras de sentimentos tristes (negativo) e sentimentos felizes (positivo). O corpus serviu de guia de seleção de termos para formação de clusters de palavras positivas e negativas para a leitura e processamento desses dados, permitindo determinar, assim, se um comentário é positivo ou negativo, baseando-se na abordagem na qual um determinado conteúdo é avaliado confrontando-o com uma base de conteúdo previamente classificado. Considerando os resultados obtidos a partir da investigação aqui relatada, pode-se destacar a sua contribuição para pesquisas no campo do processamento de lingua natural com foco em língua portuguesa, uma vez que o método proposto obteve uma taxa de precisão de 86% na classificação dos comentários submetidos ao sistema, o que pode ser indicar o potencial da abordagem proposta.

**Palavras-chave**: Emoticons, análise de sentimentos, redes sociais, Padrões de Superfície Textual, PLN.

**ABSTRACT**

With the huge amount of textual information generated day by day becomes a growing need for methods and techniques of extraction and analysis of knowledge in texts. Today there is hardly a transaction that is not made by computers and managing records that include valuable information such as trends and patterns, which if analyzed correctly, assist in decision making. Considering this context, the objective of this research was to investigate automatic analysis methods of feelings, proposing the development of a software. The tool was implemented based on a corpus of selected users comments, a priori, manually two social networks (Facebook and Twitter), containing # 's (hashtags) linked with words of sad feelings (negative) and happy feelings (positive). The corpus served as a selection guide for training terms of positive and negative words for reading and processing such data clusters, allowing thus determine if a review is positive or negative, based on the approach in which certain content is evaluated by comparing it with a base content previously classified. Considering the results obtained from the research reported here, we can highlight its contribution to research in natural language processing field with focus on English language, since the proposed method achieved a 86% accuracy rate in the classification of reviews submitted to the system, which can indicate the potential of the proposed approach.

**Keywords:** Emoticons, feelings analysis, social network, textual Patterns of Surface, PLN.

**LISTA DE FIGURAS**

[Figura 1 - Áreas e subáreas da Eng. de Software. 14](#_Toc404701991)

[Figura 2 - Fases do Ciclo de vida do projeto de MD 19](#_Toc404701992)

[Figura 3 - Evolução dos sistemas de PLN 23](#_Toc404701993)

[Figura 4 - Principais Fases de um Sistema de Geração de Língua Natural. 29](#_Toc404701994)

[Figura 5 - Esquema de uma Arquitetura de Sistema de Interpretação de Língua Natural. 32](#_Toc404701995)

[Figura 6 - Exemplo de comentário triste retirado do Facebook 36](#_Toc404701996)

[Figura 7 - Exemplo de comentário triste retirado do Twitter 36](#_Toc404701997)

[Figura 8 - Erro de formatação ao retirar comentário do Facebook 37](#_Toc404701998)

[Figura 9 - Visão geral do arquivo XML que contem o corpus. 38](#_Toc404701999)

[Figura 10 - Contagem dos termos. 39](#_Toc404702000)

[Figura 11 - Ranque de pesos dos termos do Cluster feliz. 40](#_Toc404702001)

[Figura 12 - Ranque de pesos dos termos do Cluster triste. 40](#_Toc404702002)

[Figura 13 - Quantidade de termos nos clusters feliz e triste. 41](#_Toc404702003)

[Figura 14 - Software de reconhecimento de sentimentos 43](#_Toc404702004)

[Figura 15 - Software de reconhecimento de sentimentos exportando arquivo tokensTriste.txt. 44](#_Toc404702005)

[Figura 16 - Software de reconhecimento de sentimentos exportando arquivo tokenTexto.txt 44](#_Toc404702006)

[Figura 17 - Software de reconhecimento de sentimentos reconhecendo o sentimento expresso em determinado comentário já devidamente tratado. 45](#_Toc404702007)

[Figura 18 - Gráfico de resultados do reconhecimento de comentários do Facebook com #(Hashtag) indicativa de sentimento. 46](#_Toc404702008)

[Figura 19 - Gráfico de resultados do reconhecimento de comentários do Facebook sem #(Hashtag) indicativa de sentimento. 46](#_Toc404702009)

[Figura 20 - Gráfico de resultados do reconhecimento de comentários do Twitter com #(Hashtag) indicativa de sentimento. 47](#_Toc404702010)

[Figura 21 - Gráfico de resultados do reconhecimento de comentários do Twitter sem #(Hashtag) indicativa de sentimento. 47](#_Toc404702011)

[Figura 22 - Gráfico da média simples dos resultados do reconhecimento de comentários tristes do Twitter e Facebook com e sem #(Hashtag) indicativa de sentimento. 48](#_Toc404702012)

[Figura 23 - Gráfico de resultados obtidos dos comentários felizes do Facebook com #(Hashtag) indicativa de sentimento. 48](#_Toc404702013)

[Figura 24 - Gráfico de resultados do reconhecimento de comentários felizes recuperados no Facebook sem #(Hashtag) indicativa de sentimento. 49](#_Toc404702014)

[Figura 25 - Gráfico de resultados obtidos no reconhecimento de comentários do Twitter com #(Hashtag) indicativa de sentimento. 49](#_Toc404702015)

[Figura 26 - Gráfico de resultados obtidos no reconhecimento de comentários do Twitter sem #(Hashtag) indicativa de sentimento. 50](#_Toc404702016)

[Figura 27 - Gráfico da média simples dos resultados do reconhecimento de comentários de sentimento de felicidade do Twitter e Facebook com e sem #(Hashtag) indicativa de sentimento. 50](#_Toc404702017)

[Figura 28 - Gráfico dos resultados do reconhecimento dos comentários tristes e felizes com e sem #(Hashtag) do Facebook e do Twitter. 51](#_Toc404702018)

[Figura 29 - Gráfico dos resultados do reconhecimento de todos os comentários com todas as possibilidades de teste. 51](#_Toc404702019)

**SUMÁRIO**

[1 INTRODUÇÃO 11](#_Toc404701961)

[2 OBJETIVOS 13](#_Toc404701962)

[2.1 OBJETIVO GERAL 13](#_Toc404701963)

[2.2 OBJETIVOS ESPECÍFICOS 13](#_Toc404701964)

[3 ENGENHARIA DE SOFTWARE 14](#_Toc404701965)

[3.1 O PROCESSO DE SOFTWARE 15](#_Toc404701966)

[4 INTELIGÊNCIA ARTIFICIAL 18](#_Toc404701967)

[5 MINERAÇÃO DE DADOS 19](#_Toc404701968)

[6 PROCESSAMENTO DE LÍNGUA NATURAL (PLN) 21](#_Toc404701969)

[6.1 HISTÓRIA DO PROCESSAMENTO DE LINGUAGEM NATURAL 21](#_Toc404701970)

[6.2 O QUE É PLN 23](#_Toc404701971)

[6.3 APLICAÇÕES 24](#_Toc404701972)

[6.4 QUESTÕES CONCEITUAIS DO PLN 26](#_Toc404701973)

[6.5 ARQUITETURA DE UM SPLN 27](#_Toc404701974)

[6.6 SISTEMA DE GERAÇÃO DE LÍNGUA NATURAL 28](#_Toc404701975)

[6.7 SISTEMA DE INTERPRETAÇÃO DE LÍNGUA NATURAL 31](#_Toc404701976)

[6.8 CORPUS 33](#_Toc404701977)

[6.8.1 Linguística de Corpus 33](#_Toc404701978)

[6.9 ANÁLISE DE SENTIMENTOS 34](#_Toc404701979)

[6.9.1 Fases da Mineração de Sentimentos 35](#_Toc404701980)

[7 METODOLOGIA 36](#_Toc404701981)

[7.1 MONTAGEM DO CORPUS 36](#_Toc404701982)

[7.2 FORMAÇÃO DE UM RANQUE DE POLARIDADE DE EXPRESSÕES 39](#_Toc404701983)

[7.3 CLASSIFICAÇÃO DE COMENTÁRIOS 41](#_Toc404701984)

[8 RESULTADOS 43](#_Toc404701985)

[9 CONSIDERAÇÕES FINAIS 52](#_Toc404701986)

[REFERÊNCIAS 53](#_Toc404701987)

[ANEXO A 55](#_Toc404701988)

[ANEXO B 56](#_Toc404701989)

[ANEXO C 65](#_Toc404701990)

# 1 INTRODUÇÃO

É inegável a contribuição que os computadores têm dado para o avanço nos mais diversos campos científicos, desde que ele passou a fazer parte da nossa cultura nos anos 40, auxiliando o processo de construção de conhecimentos sofisticados, pautados sobretudo na necessidade de processamento extensivo de dados. Entretanto, a realização de tarefas por essas máquinas ainda está ligada fortemente a um processo de interação homem-máquina baseado em linguagens artificiais (linguagens de programação), que se caracterizam por apresentarem uma representação rígida imposta pela arquitetura do computador sendo de certo modo, um fator limitante apesar das chamadas linguagens de alto nível que procuram ser mais inteligíveis, pelo menos do ponto de vista humano.

A internet vem aumentando exponencialmente sua quantidade de informações textuais geradas nesta ultima década: somente nos últimos 13 anos, o numero de domínios cresceu de 15.000 para 350.000.000. (O’DELL, 2011). Esse aumento de informações também se reflete no aumento da participação dos usuários nos conteúdos da WEB por meio das redes sociais, por exemplo: Facebook com 1,23 bilhão de usuários ativos (FACEBOOK, 2014), e o Twitter com mais de 230 milhões de usuários ativos por mês (TWITTER, 2014). Tanta informação agrupada em todo esse espaço comum e público, onde são criadas centenas de milhares de novos conteúdos por dia torna extremamente difícil  e demorada a análise dos mesmos, mesmo que estes sejam uma peça, hoje, fundamental para a tomada de decisão de muitas grandes corporações e governos. Nesta necessidade, surgiu uma nova área de estudo relacionado ao Processamento de Linguagem Natural, chamada de Análise de Sentimentos. As pesquisas nesta área tem como objetivos entender de forma automática, de certo modo, a consciência coletiva dos usuários da internet. Deste modo, é possível classificar textos não por tópicos, e sim pelos sentimentos ou opiniões contida em determinado documento. Geralmente associado ao tratamento computacional de um texto permitindo a classificação binária entre sentimentos positivos e negativos. (PANG; LEE, 2008).

A Análise de Sentimentos procura, portanto, identificar o sentimento que os usuários apresentam a sobre alguma entidade de interesse (uma empresa, um produto específico, um local, uma pessoa, etc.) baseado no conteúdo textual disponível na Web. Por meio desses dados, essa prática identifica a emoção das pessoas expressa de forma escrita ou falada, especialmente em mídias sociais. Conhecer essas emoções e agir com rapidez é um dos atributos da inteligência competitiva e representa o diferencial entre o sucesso ou o fracasso de um projeto, justificando os esforços envolvidos no desenvolvimento de ferramentas que possam processar esse tipo de informação. Deste modo, cada vez mais pessoas e principalmente empresas, estão interessadas em observar as opiniões de um grupo de pessoas sobre temas que lhe interessam. Uma empresa pode, por exemplo, medir a aceitação de um novo produto, monitorando as opiniões de um grupo em relação a ele. Um partido político poderia saber o que as pessoas pensam sobre determinado candidato. Por conta desta rigidez das linguagens de programação, muitas pesquisas têm sido desenvolvidas tendo como foco a transformação dos computadores em máquinas cada vez mais acessíveis. Conforme aponta Silva et al. (2007) uma alternativa para este problema é criar programas capazes de interpretar mensagens codificadas em línguas naturais, de modo que as máquinas “aprendam” essas línguas e sejam capazes de decifrá-las, tornando “mais natural” a comunicação entre o homem e a máquina.

Diante desta problemática, este trabalho visa proporcionar subsídios para se estudar os métodos necessários para fazer com que opiniões possam ser processadas e classificadas automaticamente, gerando um “entendimento” do seu conteúdo semântico. Se considerarmos o foco particular do processamento da língua portuguesa, serão relevantes as contribuições deste trabalho, não só para a validação de uma metodologia de análise de sentimentos em redes sociais, como também para as próprias tarefas de processamento de conteúdo textual, sobretudo na web, principalmente porque há muito poucos recursos para o processamento da língua portuguesa e não existem atualmente muitos sistemas específicos de processamento de documentos nessa língua.

# 2 OBJETIVOS

## 2.1 OBJETIVO GERAL

Explorar as potencialidades dos padrões de superfície de texto para o desenvolvimento de um protótipo de sistema de análise de sentimentos, usando como corpus duas redes sociais (Facebook e Twitter), contribuindo para os avanços dos estudos nesta área em língua portuguesa.

## 2.2 OBJETIVOS ESPECÍFICOS

* Realizar um levantamento bibliográfico dos métodos a serem utilizados;
* Selecionar um conjunto de termos de estudo e investigação;
* Montar um corpus linguístico[[1]](#footnote-1) (coleção de documentos) para a realização do processamento desejado, utilizando o Facebook e Twitter;
* Criar clusters (abordagem bag-of-words[[2]](#footnote-2)) de termos positivos e negativos para classificação de sentimentos;
* Programar o protótipo em linguagem Phyton com banco de dados MySQL,  que processe os documentos segundo a metodologia proposta;
* Testar e interpretar os resultados obtidos a partir do protótipo em funcionamento.

# 3 ENGENHARIA DE SOFTWARE

Pressman (2011) define em seus trabalhos com engenharia de software que existem alguns fatores contemporâneos a serem observados:

* O software se tornou uma parte intrínseca da sociedade atual, estando em todas as camadas da vida cotidiana, e o numero de pessoas que se interessam em utilizar esse recurso tem apenas aumentado, por isso quando uma aplicação é desenvolvida muitas opiniões são levadas em conta, por vezes essas opiniões são ligeiramente divergentes sobre as funções que determinado software deve conter. Portanto compreende-se que deve ter um esforço para compreender o problema antes de desenvolver a solução para o mesmo.
* Todos os setores da sociedade estão demandando cada vez mais aplicações de alta complexidade dia após dia, essa complexidade implica uma maior atenção para as interações dos elementos de um sistema, tornando assim o planejamento uma atividade essencial para o desenvolvimento de um software.
* Dependência crescente de softwares nas tomadas de decisão de empresas, governos e indivíduos, obrigando assim o software ter qualidade elevada evitando falhas.
* Quanto mais elevado é o custo de produção de uma aplicação provavelmente o mesmo será usado por uma grande base de usuários, e também por mais tempo, aumentando a demanda por adaptações o software, tornando assim ele passível de manutenção.

Através destas constatações Pressman (2011) chega a conclusão que o software, em todas as suas formas e campos de aplicação devem passar pelos processos de engenharia.

A engenharia de software por sua vez é fortemente apoiada no principio da qualidade, pois apenas assim pode promover uma cultura de aperfeiçoamento continuo dos processos, levando ao desenvolvimento de abordagens cada vez mais efetivas.

Figura 1 - Áreas e subáreas da Eng. de Software.



Fonte: Pressman (2011, p 8).

Pressman (2011) com base na figura 1 define que o pilar da engenharia de software se dá na camada de “processos”, esta é a que mantém as camadas de tecnologia e da suporte para o desenvolvimento do software dentro dos prazos estabelecidos.Por sua vez o controle de gerenciamento de software tem como base os processos de software para formar os contextos onde: são aplicados os métodos técnicos, são produzidos os produtos derivados, estabelecidos os marcos, a qualidade é garantida e mudanças são geradas da melhor maneira possível.

Os métodos são responsáveis por fornecer as tecnicalidades para o desenvolvimento do software, para isso são envoltas varias tarefas de responsabilidade dos mesmos, como: comunicação, análise de requisitos, modelagem, de projeto, construção de programa, teste de suporte. Eles baseiam-se em um conjunto de princípios básicos que gerem cada parte da tecnologia

Segundo Pressman (2011) as ferramentas da engenharia de software proporciona uma estrutura automatizada ou quase para as camadas de processo e de métodos. No momento em que as ferramentas são integradas entre si para que as informações geradas por uma e usadas na outra é estabelecido um sistema para o desenvolvimento da engenharia de software chamado engenharia de software com o auxilio do computador.

## 3.1 O PROCESSO DE SOFTWARE

Pressman (2011) define que um processo de software pode ser visto como o conjunto de atividades, métodos, práticas e transformações que guiam pessoas na produção de software. Um processo eficaz deve, claramente, considerar as relações entre as atividades, os artefatos produzidos no desenvolvimento, as ferramentas e os procedimentos necessários e a habilidade, o treinamento e a motivação do pessoal envolvido.

Há vários aspectos a serem considerados na definição de um processo de software. No centro da arquitetura de um processo de desenvolvimento estão as atividades-chave desse processo: análise e especificação de requisitos, projeto, implementação e testes, que são a base sobre a qual o processo de desenvolvimento deve ser construído. Entretanto, a definição de um processo envolve a escolha de um modelo de ciclo de vida, o detalhamento (decomposição) de suas macro-atividades, a escolha de métodos, técnicas e roteiros (procedimentos) para a sua realização e a definição de recursos e artefatos necessários e produzidos Pressman (2011).

Contudo Pressman (2011) diz que um processo de software não pode ser definido de forma universal. Para ser eficaz e conduzir à construção de produtos de boa qualidade, um processo deve ser adequado ao domínio da aplicação e ao projeto específico. Deste modo, processos devem ser definidos caso a caso, considerando-se as especificidades da aplicação, a tecnologia a ser adotada na sua construção, a organização onde o produto será desenvolvido e o grupo de desenvolvimento. Em suma, o objetivo de se definir um processo de software é favorecer a produção de sistemas de alta qualidade, atingindo as necessidades dos usuários finais, dentro de um cronograma e um orçamento previsíveis.

De maneira geral Pressman (2011) descreve o ciclo de vida de um software, que envolve as seguintes fases:

• *Planejamento*: O objetivo do planejamento de projeto é fornecer uma estrutura que possibilite ao gerente fazer estimativas razoáveis de recursos, custos e prazos. Uma vez estabelecido o escopo de software, uma proposta de desenvolvimento deve ser elaborada, isto é, um plano de projeto deve ser elaborado configurando o processo a ser utilizado no desenvolvimento de software. À medida que o projeto progride, o planejamento deve ser detalhado e atualizado regularmente. Pelo menos ao final de cada uma das fases do desenvolvimento (análise e especificação de requisitos, projeto, implementação e testes), o planejamento como um todo deve ser revisto e o planejamento da etapa seguinte deve ser detalhado. O planejamento e o acompanhamento do progresso fazem parte do processo de gerência de projeto.

• *Análise e Especificação de Requisitos*: Nesta fase, o processo de levantamento de requisitos é intensificado. O escopo deve ser refinado e os requisitos identificados. Para entender a natureza do software a ser construído, o engenheiro de software tem de compreender o domínio do problema, bem como a funcionalidade e o comportamento esperados. Uma vez identificados os requisitos do sistema a ser desenvolvido, estes devem ser modelados, avaliados e documentados. Uma parte vital desta fase é a construção de um modelo descrevendo *o que* o software tem de fazer (e não *como* fazê-lo).

• *Projeto*: Esta fase é responsável por incorporar requisitos tecnológicos aos requisitos essenciais do sistema, modelados na fase anterior e, portanto, requer que a plataforma de implementação seja conhecida. Basicamente, envolve duas grandes etapas: projeto da arquitetura do sistema e projeto detalhado. O objetivo da primeira etapa é definir a arquitetura geral do software, tendo por base o modelo construído na fase de análise de requisitos. Esta arquitetura deve descrever a estrutura de nível mais alto da aplicação e identificar seus principais componentes. O propósito do projeto detalhado é detalhar o projeto do software para cada componente identificado na etapa anterior. Os componentes de software devem ser sucessivamente refinados em níveis de maior detalhamento, até que possam ser codificados e testados.

• Implementação: O projeto deve ser traduzido para uma forma passível de execução pela máquina. A fase de implementação realiza esta tarefa, isto é, cada unidade de software do projeto detalhado é implementada.

• Testes: inclui diversos níveis de testes, a saber, teste de unidade, teste de integração e teste de sistema. Inicialmente, cada unidade de software implementada deve ser testada e os resultados documentados. A seguir, os diversos componentes devem ser integrados sucessivamente até se obter o sistema. Finalmente, o sistema como um todo deve ser testado.

• Entrega e Implantação: uma vez testado, o software deve ser colocado em produção. Para tal, contudo, é necessário treinar os usuários, configurar o ambiente de produção e, muitas vezes, converter bases de dados. O propósito desta fase é estabelecer que o software satisfaz os requisitos dos usuários. Isto é feito instalando o software e conduzindo testes de aceitação (validação). Quando o software tiver demonstrado prover as capacidades requeridas, ele pode ser aceito e a operação iniciada.

• Operação: nesta fase, o software é utilizado pelos usuários no ambiente de produção.

• Manutenção: Indubitavelmente, o software sofrerá mudanças após ter sido entregue para o usuário. Alterações ocorrerão porque erros foram encontrados, porque o software precisa ser adaptado para acomodar mudanças em seu ambiente externo, ou porque o cliente necessita de funcionalidade adicional ou aumento de desempenho. Muitas vezes, dependendo do tipo e porte da manutenção necessária, essa fase pode requerer a definição de um novo processo, onde cada uma das fases precedentes é re-aplicada no contexto de um software existente ao invés de um novo.

# 4 INTELIGÊNCIA ARTIFICIAL

Para se conceituar Inteligência Artificial (IA), leva-se em conta a interação com o ambiente, diante de necessidades reais. McCarthy (1955) na conferência de Dartmouth dá uma definição inicial de IA que, segundo ele, é "[..] fazer a máquina comportar-se de tal forma que seja chamada inteligente caso fosse este o comportamento de um ser humano."

Inteligência Artificial é uma área de pesquisa da [computação](http://pt.wikipedia.org/wiki/Computa%C3%A7%C3%A3o) aplicada na busca de métodos ou dispositivos que contenham ou multipliquem a capacidade racional do homem de encontrar a solução de problemas, pensar ou, de alguma forma ser inteligente. Pode-se defini-la como o vertente da computação que tem a função da conduta inteligente. (LUGER, 2004).

O conceito de Inteligência Artificial está atrelado à duas questões: o que é artificial, e o que é inteligência. A primeira questão é fácil de ser delimitada, pois o artificial é tudo aquilo que pode ser criado de formas não naturais, ou também aquilo que é criado pela ação direta do homem.

A segunda é um pouco mais difícil de ser definida, pois seria levada em questão a [consciência](http://pt.wikipedia.org/wiki/Consci%C3%AAncia), [identidade](http://pt.wikipedia.org/wiki/Identidade) e [mente](http://pt.wikipedia.org/wiki/Mente) juntamente com a questão de que componentes estão envolvidos no único tipo de inteligência que está ao alcance do nosso estudo: a inteligência do ser [humano](http://pt.wikipedia.org/wiki/Humano).

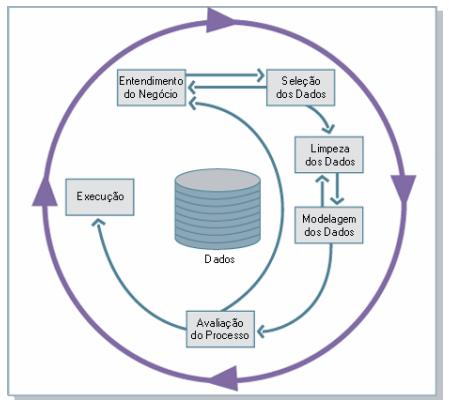
**5 MINERAÇÃO DE DADOS**

Amorin (2006) diz que a mineração de dados pode ser definida como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano.

Segundo Amorin (2006) diz que o atual processo para mineração de dados propõe uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases correspondentes do projeto, suas respectivas tarefas e relacionamentos entre essas tarefas.

Amorin (2006) relata na figura 2 que o ciclo de vida de um projeto de MD consiste em 6 fases, mas a seqüência de fases não é obrigatória, ocorrendo a transição para diferentes fases, dependendo do resultado de cada fase, e que etapa particular de cada fase precisa ser executada em seguida. As setas indicam as mais importantes e mais freqüentes dependências entre as fases.

Figura 2 - Fases do Ciclo de vida do projeto de MD



Fonte: Amorin (2006, pg. 19)

Segue abaixo uma síntese das etapas pertencentes ao modelo representado na figura 2 segundo Amorin (2006):

* Entendimento do Negócio: Essa fase inicial tem o foco no entendimento do negócio que visa obter conhecimento sobre os objetivos do negócio e seus requisitos, e então converter esse conhecimento em uma definição de um problema de mineração de dados, e um plano preliminar designado para alcançar esses objetivos;
* Seleção dos Dados: Consiste no entendimento dos dados, que visa à familiarização com o banco de dados pelo grupo de projeto, utilizando-se de conjuntos de dados "modelo". Uma vez definido o domínio sobre o qual se pretende executar o processo de descoberta, o próximo passo é selecionar e coletar o conjunto de dados ou variáveis necessárias. Inicia-se com uma coleta inicial de dados, e com procedimentos e atividades visando a familiarização com os dados, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formar hipóteses.
* Limpeza dos Dados: A fase de preparação de dados consiste na preparação dos dados que visa à limpeza, integração e formatação dos dados da etapa anterior. É a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados. Esta fase abrange todas as atividades para construir o conjunto de dados final (dados que serão alimentados nas ferramentas de mineração), a partir do conjunto de dados inicial.
* Modelagem dos Dados: Fase que consiste na modelagem dos dados, a qual visa a aplicação de técnicas de modelagem sobre o conjunto de dados preparado na etapa anterior. Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para se obter valores otimizados. Geralmente, existem várias técnicas para o mesmo tipo de problema de mineração. Algumas técnicas possuem requerimentos específicos na forma dos dados. Conseqüentemente, voltar para a etapa de preparação de dados é frequentemente necessário. A maioria das técnicas de mineração de dados é baseada em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação e clusterização.
* Avaliação do processo: A avaliação do processo visa garantir que o modelo gerado atenda às expectativas da organização. Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer um dos estágios anteriores do processo de mineração.
* Execução: Esta fase consiste na definição das fases de implantação do projeto de Mineração de Dados. A criação do modelo não é o fim do projeto. Mesmo se a finalidade do modelo for apenas aumentar o conhecimento dos dados, o conhecimento ganho necessitará ser organizado e apresentado em uma maneira que o cliente possa usar. Dependendo das exigências, a fase de execução pode ser tão simples quanto a geração de um relatório, ou tão complexo quanto executar processos de mineração de dados repetidamente.

# 6 PROCESSAMENTO DE LÍNGUA NATURAL (PLN)

Segundo Silva et al. (2007) a linguagem natural é uma linguagem utilizada para comunicação entre seres humanos, nascida de forma natural, de modo não premeditado, e o seu estudo é chamado de linguística. A linguagem é tanto a capacidade humana para a utilização de formas de comunicação complexas, quanto um sistema específico de comunicação. O Processamento de Língua Natural (PLN)  é uma subárea da Inteligência Artificial e da Linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais.

## 6.1 HISTÓRIA DO PROCESSAMENTO DE LINGUAGEM NATURAL

O início das pesquisas com PLN, de acordo com Silva et al. (2007), se iniciaram nos anos 50, após a distribuição de 200 cópias da carta Weaver Memorandum, redigida pelo então vice-presidente da Fundação Rockfler e um grande entusiasta da criptografia computacional. Continha na carta um convite de Weaver as universidades e corporações para o desenvolvimento de projetos em um novo campo, campo esse que ficou conhecido “tradução automática” ou MT (“Machine Translation”).

Este documento que foi divulgado em 1949 continha as primeiras teorias e métodos sobre alguns pontos fundamentais que deveriam ser levados em conta nesse tipo de estudo, mas suas diretrizes não estavam no centro das pesquisas dos projetistas da época. Neste tempo, traduzir um texto ou frase não se diferenciava de descriptografar algo, esta técnica se difere em muito no processo de tradução, tornando-a completamente inadequada ao tratamento de linguagens naturais, entretanto era a única arma que tinham para desenvolver programas tradutores. (SILVA et al., 2007).

Conforme Silva et al. (2007) universidades como o MIT, Harvard, Califórnia e Georgetown passaram a levar a iniciativa à frente, levantando vários pontos que deveriam ser pesquisados. A primeira reunião envolvendo o tema ocorreu no MIT em 1952 e a primeira demonstração ao público ocorreu em 54 na Universidade de Georgetown. Nesta demonstração foi apresentado um sistema capaz de traduzir um texto de química com cerca de 50 frases. A tradução bem “sucedida” desta primeira apresentação ocasionou um grande alvoroço em cima dos estudos sobre PLN, atraindo a atenção de muitas instituições financiadoras de vários países e incluindo a antiga União Soviética.

O dicionário deste sistema continha apenas 250 palavras e seis regras gramaticais de russo. Para abranger outras estruturas léxicas e mais regras gramaticais de uma quantidade maior de línguas ocorreram muitas outras tentativas de reproduzir este experimento. Mas os resultados obtidos ficaram muito abaixo das expectativas das intuições financiadores.

Bar-Hillel o maior crítico de sistemas de tradução automática afirma que uma tradução completamente automática de qualidade era impossível, pois esses sistemas listavam as possibilidades de tradução de cada palavra sem análise gramatical, logo essas traduções por conta da péssima qualidade precisavam de revisões constantes dos pesquisadores.

Em 1964, foi elaborado e divulgado um relatório do Comitê Assessor de Processamento Automático das Línguas Naturais com uma avaliação negativa sobre todas as pesquisas já realizadas, e também dizia que não havia possibilidade de futuro na tradução automática sem a necessidade de tradutores humanos revisando todos os textos traduzidos de forma automática. Por conta de toda essa repercussão dos meios especializados as pesquisas com tradução automática sofreram uma perda de recursos substancial dos financiadores, com isso o ritmo das pesquisas perdeu força. (SILVA et al., 2007).

Depois de muitas pesquisas fracassadas e ou de pouco crédito, como por exemplo, contar quantas palavras “king” existia em obras de Shakespeare, toda a área de PLN sofreu um revés, pois eles não mostravam fundamentação gramatical e linguística.

O relatório da ALPAC prejudicou em muito muitos projetos que caminhavam para o sucesso, projetos esses que levavam em conta regras gramaticais mais elaboradas e teoria lingüística, como por exemplo, o protótipo do sistema GAT de 1962 desenvolvido na Universidade de Georgetown que era capaz de traduzir com uma qualidade interessante textos do inglês para o russo. Apenas nos anos 70 os projetos em PLN voltaram com mais força por conta dos avanços tecnológicos de sua época.

Silva et al. (2007), citam alguns projetos mais maduros na área: TAUM-METEO, SYSTRAN, ATLAS II, EUROTRA e KBMT, desenvolvidos entre os anos 70 e 80 logo após a separação dos projetos e seus pesquisadores dos interesses da guerra fria. Nesta época os estudos passaram a ser mais minuciosos perante a alta complexidade do processo de tradução e do código linguístico sofisticado.

Em 1970 o estudante de doutorado Winograd do MIT, em sua tese criou um sistema que ficou como referência para os estudos de PLN e tradução automática. O SHRDLU simulava um braço robótico por meio de representação gráfica, este braço manipulava blocos sobre uma mesa através da interação humana por comandos digitados em inglês, mostrando assim para toda comunidade que a interação homem - maquina poderia ser uma realidade por meio de uma linguagem natural mesmo de forma primitiva (Silva et al.2007),.

Através de trabalhos como esse, o Processamento de Linguagens Naturais passou de fato a ser pesquisado não somente meio acadêmico, mas nade forma comercial à um passo extraordinário. Na figura 3 seguinte pode-se ver toda a evolução dos estudos em PLN de forma resumida em termos de complexidade gramatical e linguística:

Figura 3 - Evolução dos sistemas de PLN

|  |
| --- |
| Década de 50: A Tradução automática   * Sistematização computacional das classes de palavras da gramática tradicional * Identificação computacional de poucos tipos de constituintes oracionais   Década de 60: Novas aplicações e criação de formalismos   * Primeiros tratamentos computacionais das gramáticas livres de contexto * Criação dos primeiros analisadores sintáticos * Primeiras formalizações do significado em termos de redes semânticas   Década de 70: Consolidação dos estudos do PLN   * Implementação de parcelas das primeiras gramáticas e analisadores sintáticos * Busca de formalização de fatores pragmáticos e discursivos   Década de 80: Sofisticação dos sistemas   * Desenvolvimento de teorias lingüísticas motivadas pelos estudos do PLN   Década de 90: Sistemas baseados em “representações do conhecimento”   * Desenvolvimento de projetos de sistemas de PLN complexos que buscam a integração dos vários tipos de conhecimentos lingüísticos e extralingüísticos e das estratégias de inferência envolvidos nos processos de produção, manipulação e interpretação de objetos lingüísticos |

Fonte: Silva et al. (2007, p 8).

## 6.2 O QUE É PLN

Segundo Gozales, M.; Lima, V.L.S. (c2007-2008), o PLN (Processamento de Linguagem Natural) tem por objetivo os mais diversos aspectos da comunicação humana, como discursos, sentenças, palavras, e sons, levando em com ta seus usos, contextos, estruturas e significados, e formatos e referências. Ou seja, o PLN tende a fazer o computador entender e se comunicar em linguagem humana, mas não necessariamente em todos os níveis de entendimento, que são:

* Pragmático: uso de sentenças e fazes em diferentes contextos, assim alterando o seu significado.

Sintático: regras e princípios que organizam a construção das frases e sentenças através das palavras e seus significados (SINTAXE, c2013);

* Semântico: significado das palavras e como elas são usadas na construção de frases e sentenças;
* Morfológico: da formação e classificação morfológica das palavras a partir de unidades primitivas;
* Fonológico: sons que são produzidos pelas palavras.

O significado de uma sentença pode ser obtido, segundo Gozales, M.; Lima, V.L.S. (c2007-2008), de sua forma lógica. A forma lógica identifica a semântica na frase codificando os possíveis sentidos de cada palavra, uma vez prontos esses relacionamentos semânticos podemos desconsiderar alguns sentidos de palavras, pois assim tornam-se inviáveis.

A sintática de uma frase é gerada através do processamento morfosintático, sendo regidas pelas leis gramaticais que pertencem a uma gramática. Mais informações como as categorias morfológicas das palavras podem ser encontradas em um léxico.

## 6.3 APLICAÇÕES

Serão listados a seguir os principais tipos de aplicações em PLN, como descreve Silva et al. (2007), pois existem inúmeros graus diferentes de sofisticação linguística tornando, assim, as aplicações em PLN de modo geral expressivas e impressivas.

* Manipulação de bases de dados: O PNL neste tipo de sistema é de comunicação entre o usuário e o banco de dados traduzindo comandos dados pelo usuário, comandos que depois eram codificados em linguagem especifica do banco de dados que se encarregava das manipulações das respostas. Este tipo de PLN é nominado de Sistema de Perguntas e Respostas.
* Sistema tutor: É um sistema que sem a interferência humana dá aos alunos feedbacks diretos e personalizados. São feitos para simular um tutor humano simulando assim todo um diálogo levando em conta uma rede de conhecimento e fatos, regras e relações. Esse tipo de sistema se destaca por que tem uma riqueza de detalhes nas pesquisas que geram assim permitindo ao desenvolvedor fazer vários tipos de simulações.
* Sistemas de automação de tarefas administrativas: Esse tipo de sistema facilita as tarefas administrativas dos setores gerenciais de empresas, eles podem detectar erros ortográficos e gramaticais analisando palavras e frases em textos de documentos administrativos, pode também manipular objetos por meio de linguagem oral, gerenciar agendas e planejar informações de viagens aéreas, entre outras rotinas administrativas.
* Programação automática: Neste, os sistemas tem o fim de facilitar as atividades dos programadores. Sua estrutura é muito complexa devido à enorme quantidade de tarefas, como: receber e organizar informações dadas pelo programador, devolver os elementos necessários para a programação, coordenar a síntese necessária dos programas gerados e gerar um sistema. Para realizar todas essas tarefas o sistema necessita de um modelo dos processos que deve executar, faz todas as correções necessárias, seleciona as estruturas mais apropriadas e as executa.
* Sistemas de processamento de textos científicos: Este sistema recebe as informações de entrada do usuário em forma de perguntas e as codifica em frases que são analisadas em um padrão pré-existente definido a partir de regras gramaticais e sintáticas das palavras, depois é armazenado em forma de tabela que contem todas as informações da frase pergunta do usuário e da frase resposta que o sistema deve fornecer.
* Tradução automática: A tradução automática pode ser dividida em três tipos de sistemas de acordo com sua metodologia:
  + Sistemas diretos: são sistemas que tem por prioridade a correlação direta das unidades léxicas da língua a ser traduzida e língua qual é o objetivo da tradução.
  + Sistemas transferênciais: Mais sofisticados que os diretos, eles constroem as regras de sintática da língua traduzida através de um estudo da sintática da língua tradutora construindo uma representação sintática.
  + Sistemas interlínguais: Ele constrói uma interlíngua no meio das duas línguas, gerando uma abstração do significado da língua de partida para assim traduzir e gerar a língua de chegada. O mais complexo dos sistemas de tradução automática.

## 6.4 QUESTÕES CONCEITUAIS DO PLN

Winograd (1972) deixou um norte para pesquisas com PLN e pode-se visualizar isto em alguns textos deixados por ele, como por exemplo:

Assumimos que um computador não poderá simular uma língua natural satisfatoriamente se não compreender o assunto que está em discussão. Logo, é preciso fornecer ao programa um modelo detalhado do domínio específico do discurso. Além disso, o sistema possui um modelo simples de sua própria mentalidade. Ele pode se lembrar de seus planos e ações, discuti-los e executá-los. Ele participa de um diálogo, respondendo, com ações e frases, às frases digitadas em inglês pelo usuário; solicita esclarecimentos quando seus programas heurísticos não conseguem compreender uma frase com a ajuda das informações sintáticas, semânticas, contextuais e do conhecimento de mundo físico representadas dentro do sistema. (WINOGRAD, 1972).

Pode-se verificar com a concepção de PLN de Winograd (1972) que para simular uma língua natural precisam-se conter vários sistemas de conhecimento e ser capaz de realizar varias atividades cognitivas, tais como:

* Possuir um “modelo simples de sua própria mentalidade”;
* possuir um “modelo detalhado do domínio específico do discurso”;
* possuir um modelo que represente “informações morfológicas, sintáticas, semânticas, contextuais e do conhecimento de mundo físico”;
* “compreender o assunto que está em discussão”;
* “lembrar, discutir, executar seus planos e ações”;
* participar de um diálogo, respondendo, com ações e frases, às frases digitadas pelo usuário;
* solicitar esclarecimentos quando seus programas heurísticos não conseguirem compreender uma frase.

Assim Silva et al. (2007) baseado em Winograd (1972) entende que um Sistema de Processamento de Línguas Naturais pode ter como especialidades: fazer revisões ortográficas, análises sintáticas, traduzir frases ou textos, auxiliar nas pesquisas de modelos linguísticos e fazer perguntas e respostas.

De uma forma parecida a um sistema de conhecimento a montagem de um SPLN exige três etapas para o seu desenvolvimento: a “extração do solo” que é a explicitação dos conhecimentos e habilidades linguísticas do sistema, a “lapidação” cujo é a representação formal desses conhecimentos e habilidades e a “incrustação” que por sua vez é o programa de computador que codifica essa representação.

Essas três etapas podem ser repensadas como outras três fases consecutivas da construção de parte de SPLN ou dele por completo, conforme abaixo:

* Fase Linguística: Nesta fase é montada a análise dos fenômenos linguísticos, ou seja, é realizada a construção do conhecimento sobre a própria linguagem levando em conta também os fenômenos linguísticos necessários para a construção do sistema.
* Fase Representacional: Aqui se desenvolve a parte conceitual do sistema, onde engloba a seleção de sistemas formais de representação onde se tem os resultados propostos pela fase anterior, aqui também projetam às representações lingüísticas e também as extralinguísticas.
* Fase de Implementação: Aqui nesta fase é onde acontece o desenvolvimento do sistema de forma pratica, selecionando as representações desenvolvidas nas fases anteriores e escrevendo-as em linguagens de programação, e também se estudam as questões de integração física e conceitual de todos as varias partes envolvidas.

Deste modo Silva et al. (2007) entende que desenvolver um SPLN tem por sua essência especificar, representar e implementar uma quantidade considerável de informações linguísticas e extralinguísticas, métodos de inferências e controle das mesmas, e desenvolver um sistema computacional incluindo seu software e seu hardware (quando necessário). Ou seja, desenvolver um grande programa de computador com um desempenho linguístico e metalinguístico artificial.

Sendo, a meta das pesquisas é alcançar a capacidade de projetar e programar sistemas que a comunicação homem-máquina possa ser realizada por meio da linguagem natural, e não através de comandos decodificados por uma língua artificial.

## 6.5 ARQUITETURA DE UM SPLN

Silva et al. (2007) citam como exemplo de arquitetura de SPLN os sistemas de tradução automática, pois eles podem ser considerados um dos sistemas de PLN mais completos e complexos, podendo existir numerosas possibilidades de arquitetura variando nas especificações necessárias da aplicação.

Um sistema de tradução automática deverá conter as seguintes funções segundo Silva et al. (2007):

* Reconhecer e extrair cada palavra do texto em língua original;
* analisar a sintática do texto ou frase, associando cada palavra aos seus devidos atributos;
* representar o texto analisado de uma maneira intermediaria que agrega as informações sintáticas.
* analisar a semântica do texto a ser traduzido, extraindo uma semântica global, partindo das relações entre palavras ou grupos, e seus significados;
* Associar o significado em uma representação adequada, representação essa que pode ser Independiente da língua destino;
* Transformar a representação criada anteriormente em uma sentença para língua destino.

Dos processos acima citados, dois deles, mesmo de forma isolada são muito complexos: a fase de interpretação, que consiste na transformação do texto da língua de origem em uma forma intermediaria, cujo são os quatro primeiros processos, e a fase de geração do texto em língua destino, a partir da forma intermediaria criada nos processos anteriores.

A maioria dos SPLN possui apenas uma dessas fases (interpretação, geração). Os próximos tópicos abordarão o tipo de arquitetura e suas peculiaridades de cada uma dessas fases, dos sistemas de interpretação e geração de língua natural.

## 6.6 SISTEMA DE GERAÇÃO DE LÍNGUA NATURAL

Conforme apresentado na Figura 4, pode-se visualizar uma arquitetura de um sistema automático de geração de língua natural. Nele, durante sua geração se considera três processos fundamentais: Seleção, Planejamento textual, e a Realização textual. Como podemos ver na Figura 4 a geração textual usa recursos parecidos com os da interpretação, pois o PLN tem o mesmo conhecimento linguístico ou extralinguístico independente do método usado ou do grau de profundidade do uso, o processamento da informação é distinto a ponto de não permitir a inversão dos módulos de um processo ao outro.

Figura 4 - Principais Fases de um Sistema de Geração de Língua Natural.



Fonte: Silva et al. (2007, p 30).

Sendo assim, na maior parte das vezes não podemos considerar a geração como um inverso da interpretação. O gerador tem por função produzir textos em língua natural de um agrupamento de elementos de objetivos de comunicação e conteúdo. Apesar desses métodos, em muitos PLNs a geração de língua natural é produzida de uma forma bem simples, os textos são instituídos pela justaposição de segmentos textuais já definidos durante a fase de projeto, ocasionalmente, os esquemas de texto são preenchidos para compor o texto final, em tal caso os esquemas também são prefixados, mas durante o processamento também podem possuir uma parte variável. Toda via esses métodos são satisfatórios apesar das limitações de suas técnicas.

Todas as funções citadas por Silva et al. (2007), têm o controle sob a organização do conteúdo selecionado e suas varias formas lingüísticas, que são tarefas de um produtor (humano) de textos. Para produzir à estrutura textual e seus elementos lingüísticos a gramática de geração procede a partir dos elementos conceituais do texto. Sendo assim, as decisões a respeito dos constituintes sintáticos, do vocabulário, e da forma da sentença são de completa responsabilidade do gerador a partir do momento em que é determinado como combinar os conceitos do discurso para assim atingir a comunicação pretendida.

Neste momento será mostrada a especificação funcional de cada uma das fases segundo Silva et al. (2007), para que se tenha uma representação completa do discurso para se obter a representação superficial.

* Seleção de conteúdo: Tem por função a seleção dos itens que farão parte do texto, logo no processo comunicativo está fase determina “o que será dito”;
* Planejamento do texto: Aqui é decidido “quando se dizer” o que foi selecionado anteriormente. Essa fase é chamada de componente estratégico pois é responsável por planejar a comunicação. Para ter uma melhor apresentação textual nesta faze todo o conteúdo deve ser organizado, implicando na soma de especificações retóricas, na sequência de apresentação das informações, e em decisões sobre a escolha das palavras a serem usadas. Deste modo o planejador constrói uma espécie de texto intermediário chamando de “plano do texto”. Existe uma diferenciação em relação aos formalismos de representação do plano do texto, essa diferença ocorre quando há uma necessidade de privilegiar algum aspecto do plano dependendo da aplicação, alguns privilegiam a estrutura retórica do texto e outros os aspectos pragmáticos.
* Realização: Chamado também de componente tático, é essencial para que seja possível a confecção gramatical do projeto de texto construído pelo planejador. Podemos ter dois sub-processos nesta fase: a linearização da estrutura textual seqüencialização dos elementos na forma textual, gerando uma seqüência de sentenças validas, e a determinação dos elementos linguísticos. A contribuição desse componente para o processo engloba algumas decisões linguísticas e sobre o conhecimento do discurso e do domínio tais como:
  + Definição de vocabulário;
  + Definição do estilo do texto;
  + Definições léxicas, morfológicas e sintáticas para expressar o conteúdo textual;
  + Definição de figuras de discurso para manifestar apropriadamente as intenções do autor;
  + Definição que garantam a coesão do discurso, e a fluidez do texto;
  + Definição que garantam a coerência do discurso, e que expressem o inter-relacionamento semântico desejado;
  + E decisões de linearização, p.ex., ordenação das informações, concordância gramatical, etc.

O realizador como mostrado possui como funções, além da fase da escolha das palavras:

* Esquematizar a estrutura temática de uma estrutura sintática de superfície;
* Impor a gramática, tais como concordâncias entre verbo e sujeito, substantivo e determinante;
* Selecionar pronomes, conjunções, artigos, entre outras palavras de classe fechada;
* Apresentar formas diferentes de acordo com a categoria gramatical das palavras de classe aberta;
* Tornar linear a árvore sintática em uma cadeia de palavras flexionadas.

Nos relatórios de Silva et al.,(2007) a distinção das fases de realização e planejamento traz muitas vantagens, pois dá ao menos dois níveis de abstração, para que os detalhes sejam levados em conta para o realizador e ser ignorados pelo planejador.

Do mesmo modo que na interpretação, Silva et al.,(2007) diz que para os processos de um gerador automático podemos considerar três fases:

* Sequencial: Os processos são unicamente sequenciais sendo que a realização ocorre quando o planejamento já foi finalizado, assim cada módulo não se interfere;
* Intercalado: Deferente do sequencial onde os módulos não se interferem. Neste os módulos são executados de modo intercalado, comunicando-se entre si para que quando um modulo tenha a necessidade de informações que não são de sua responsabilidade execute outro modulo;
* Combinado: São executadas todas as tarefas do processo sem a necessidade de se separar em módulos.

## 6.7 SISTEMA DE INTERPRETAÇÃO DE LÍNGUA NATURAL

Conforme se pode ver na Figura 5 há uma arquitetura geral de um Sistema de Interpretação de Língua Natural onde os recursos do processamento, gramática, léxico, ou os modelos do usuário e do domínio estão representados por elipses, e os módulos de processamento por retângulos.

Figura 5 - Esquema de uma Arquitetura de Sistema de Interpretação de Língua Natural.



Fonte: Silva et al. (2007, p.25).

Em aplicações que não precisam da interpretação de uma sentença as arquiteturas podem ser mais simplificadas, excluindo algumas bases de conhecimento ou módulos que estão representados na arquitetura da Figura 5. Mas existem outras aplicações que precisam de algum processo adicional que não está nesta arquitetura.

Silva et al. (2007) dão um resumo de cada processo e sua função da arquitetura de um Sistema de Interpretação de Linguagem Natural como sendo:

* Scanner ou Analisador Léxico: Engloba a identificação e separação dos componentes de relevância do texto sob análise, chamados de tokens;
* Parser ou Analisador Sintático: É responsável por recuperar uma estrutura sintática eficaz para o texto de entrada, para isso é levado por uma representação da gramática da língua inicial. Para evitar o enorme volume de informações gramaticais que pode aumentar a complexidade da representação e de todo processo normalmente usa-se uma gramática parcial da língua natural, apesar de não conter toda a construção da língua, têm-se todas as regras gramaticais necessárias para a aplicação.
* Analisador semântico: Responsável pela interpretação de partes da sentença ou de toda a sentença, e está sempre presente quando a aplicação exigir algum tipo de interpretação. Caso necessário, essa interpretação é preciso um conhecimento mais especifico do domínio. Enquanto a estrutura profunda de uma sentença mostra apenas a caracterização linguística e a ordem, a estrutura semântica dá o inter-relacionamento dos componentes da sentença e seus significados, sendo representadas funcionalmente a partir das agregações entre os componentes semânticos apresentados pelos componentes sentenciais.
* Analisador do Discurso: Em textos multi-senteciais, o significado de uma sentença depende das sentenças que a precede e assim pode induzir os significados das sentenças que a sucedem. Normalmente nos textos multi-senteciais são aplicados recursos lingüísticos que torna a resolução analítica mais profunda.
* Analisador Pragmático: A despeito de muitos níveis de análise da estrutura superficial do texto admitir a aquisição de uma representação semântica, a aquisição do cerne de um texto pode ser ainda estar passível a aspectos pragmáticos.

## 6.8 CORPUS

Sardinha, T.B. (2004) dá uma definição de corpus:

A Linguística de Corpus é um campo que se dedica à criação e análise de corpora (plural latim de corpus), que conjuntos de textos e transcrições de fala armazenadas em arquivos de computador. A Linguística de Corpus vem revolucionando a maneira como se investiga a linguagem, nos seus mais diversos níveis, colocando à disposição do analista quantidades de dados antes inacessíveis. Um dos grandes agentes dessa revolução é a informática; sem ela, a Linguística de Corpus contemporânea não poderia existir. Assim, o linguista de corpus depende de programas de computador para lidar com corpora. (SARDINHA, T.B., 2004).

Entende-se que exista o Corpus e seu estudo, a Linguística de Corpus, e identifica-se também o agente que vem auxiliando o desenvolvimento desses campos. Serão definidos agora os conceitos simplificados de Corpus e sua Linguística.

### 6.8.1 Linguística de Corpus

Como explicado por Sardinha, T.B. (2004) a linguística de Corpus nasceu da necessidade dos linguistas, quando houve uma necessidade de se sustentar em usos reais para criação de teorias sobre o funcionamento da língua. Hoje a linguística de corpus tem por atributos a coleta e análise de corpora digitais através de processos e métodos digitais, pois os corpora (plural de corpus) estão essencialmente ligados aos meios digitais, pois é aonde são armazenados.

A autenticidade do corpus deve ser máxima segundo Sardinha, T.B. (2006), ele não pode ser constituído de informações inventadas, essas informações devem ser legíveis por um computador e ser representativos em uma língua pela qual se tem interesse de seu estudo.

Existem vários tipos de ferramentas na análise do corpus, elas são utilizadas para a organização e extração de informação para fornecer novos pontos de vistas e perspectivas da análise da língua estudada. As mais comuns ferramentas para análise do corpus são:

* Etiquetadores: Analisam automaticamente todo o corpus e “etiquetam” cada palavra com sua semântica, gramática e morfosemântica;
* Concordanciadores: Uma ferramenta de busca utilizada por um usuário para procurar palavras específicas e mostrá-las com suas concordâncias em determinados contextos.

Toda a forma linguística não é gerada de forma aleatória podendo ser evidenciado alguns padrões, sendo comum a afirmação que a linguagem é padronizada, tendo assim relação entre os contextos situacionais do de uso da linguagem e os traços linguísticos. No corpus essa padronização é claramente vista por estruturas, coligações e colocações, que tem uma significativa repetição.

As principais áreas da Linguística de Corpus são:

* Compilação de corpora;
* Desenvolvimento de ferramentas para análise do corpus;
* Descrição de linguagem;
* Obtenção do uso de definições fundamentado em corpora para várias aplicações tal como ensino-aprendizagem de línguas, processamento de linguagem natural, reconhecimento de voz e tradução.

## 6.9 ANÁLISE DE SENTIMENTOS

Segundo Rodruies, C.A.S et al (2013), a Análise de Sentimentos tem o objetivo de identificar o sentimento que os usuários apresentam baseado no conteúdo disponível na Web. Cada vez mais pessoas e predominantemente empresas, estão interessadas em acompanhar as opiniões de um grupo de pessoas sobre elas mesmas ou seus produtos, e com essas informações tomar decisões gerenciais para mudança ou permanência de estratégias. A análise de sentimento é um tipo de mineração de dados que possibilita tal demanda.

### 6.9.1 Fases da Mineração de Sentimentos

Para realizar a mineração de sentimentos Rodruies, C.A.S et al (2013), indicam três fases necessárias:

* Coleta de Conteúdo: faz uma busca nas fontes de dados a qual se quer obter as informações;
* Classificação: Essa fase é a mais importe do processo, é nela que analisamos a polaridade do sentimento ou sua orientação, determinando assim se o sentimento é positivo, negativo ou neutro. Para se conseguir a classificação podem-se usar vários tipos de técnicas, algumas delas são:
  + Aprendizagem de máquina: Nesta fase da classificação são utilizados algoritmos conhecidos e consolidados,
  + Seleção das Palavras: Nesta técnica existem palavras previamente classificadas para assim identificar o resto do conteúdo com base na avaliação do conjunto já classificado.
  + Análise sintática: Analisa-se aqui a sintática do conteúdo a procura de seus adjetivos e ou advérbios que indicam a sua polaridade.
* Sumarização dos Resultados: Fase responsável pela apresentação dos resultados obtidos em sua forma textual e ou gráfica, com todas as informações necessárias.

# 7 METODOLOGIA

Inicialmente foi realizada um levantamento bibliográfico, visando definir os materiais e os métodos mais adequados ao desenvolvimento da ferramenta de análise de sentimentos com as características propostas nesta investigação. Uma vez concluída esta pesquisa, foram realizadas basicamente três etapas, como segue:

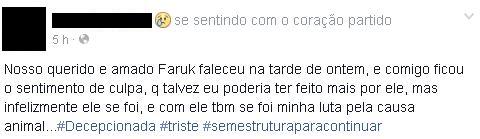
* Montagem do corpus;
* Ranking de polaridade de palavras/expressões;
* Classificação de comentários.

Tais etapas estão detalhadas nas seções seguintes.

## 7.1 MONTAGEM DO CORPUS

Etapa na qual foi feita um busca no Facebook e Twitter, nos comentários de usuários. Selecionando manualmente um conjunto de comentários por meio da ferramenta de busca. Foram recuperados para formação do corpus um total de 102 comentários marcados com # (hastag) vinculados às palavras que indiquem sentimentos de tristeza (#triste) e a sentimentos de felicidade (#feliz). A figura 6 mostra um exemplo de comentário com sentimento triste retirado do Facebook, e a figura 7 mostra um retirado do Twitter, exatamente com as #(Hashtag) que indica o sentimento do comentário, outras #s(Hashtags) foram adicionadas ao corpus mas somente as duas citadas acima serviram de norte para os testes.

Figura 6 - Exemplo de comentário triste retirado do Facebook



Fonte: Facebook (2014).

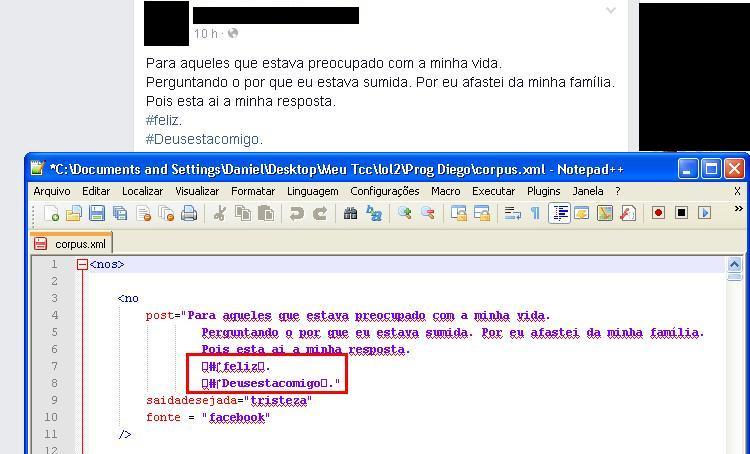
Figura 7 - Exemplo de comentário triste retirado do Twitter



Fonte: Twitter (2014).

O tamanho reduzido do corpus se deve a vários fatores, entre eles: a dificuldade da coleta (manual), e erros de formatação apresentada no momento da retirada dos comentários do Facebook como a figura 8 exemplifica.

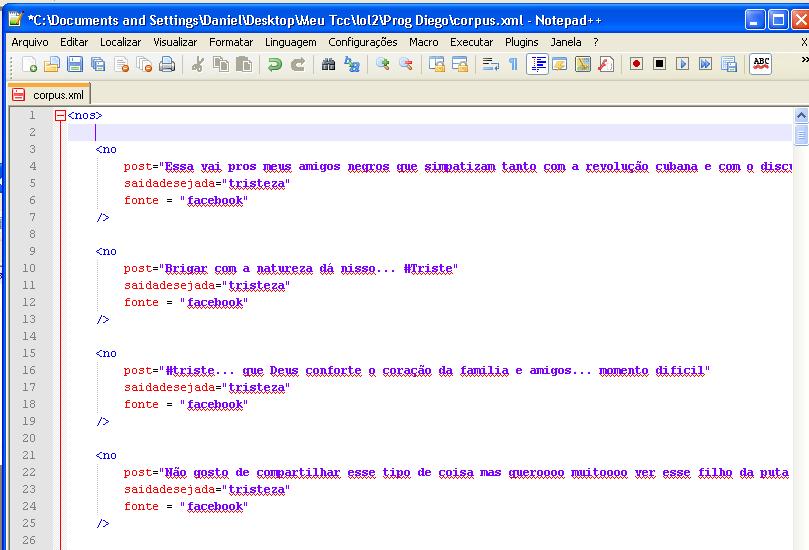
Figura 8 - Erro de formatação ao retirar comentário do Facebook



Fonte: Facebook (2014).

Os comentários recuperados foram usados para guiar a seleção de outros termos (palavras ou expressões) que formarão os clusters de palavras positivas e negativas. Todos os comentários recuperados foram armazenados em arquivos do tipo XML divididos em nós, e cada nó contendo como atributos: o post, a fonte de onde foi retirado e a saída desejada indicando qual sentimento este comentário estava representando, conforme mostra a figura 9. O atributo post contem o comentário recuperado e pré-tratado onde foi retirado os erros de formatação inerentes ao ato de recuperar tal comentário do Facebook, o atributo fonte aloca a fonte propriamente dita do comentário, Facebook ou Twitter para uma posterior classificação dos resultados e, por fim, o atributo saída desejada é responsável por ser o indicador na hora de separar o corpus único em clusters, essa saída desejada foi indicado por um juiz humano, mas poderia, por exemplo, ser atribuída automaticamente através da #(Hashtag) indicativa de sentimento.

Figura 9 - Visão geral do arquivo XML que contem o corpus.

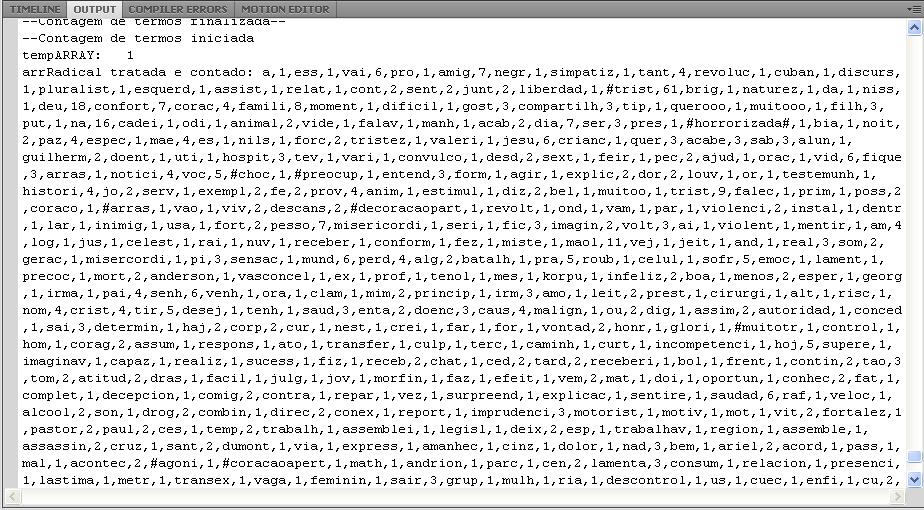


Fonte: Elaborado pelo autor.

Para criar os clusters foram necessárias três etapas: leitura e conversão, extração e limpeza dos termos e contagem de termos.

* Leitura: Nessa etapa foi feita a leitura do arquivo XML por de um método implementado em classe no software desenvolvido através do atributo “Saída Desejada”, separando, assim, os comentários que indicavam o sentimento de tristeza e o de felicidade em dois Arrays correspondentes (corpusFeliz e corpusTriste) para, assim, ter seu conteúdo textual extraído e seguir pelas etapas seguintes.
* Extração e Limpeza dos termos: Cada documento da coleção tem o seu conteúdo dividido em termos, ou seja, cada palavra/expressão significante presente no documento. Este passo é composto por 3 sub-etapas:
  + Limpeza: Primeiro são removidas todas as pontuações, depois stop words, lista no Anexo A, e, por fim, são trocados as letras com acentuação por suas correspondentes sem a mesma. A pontuação e a acentuação são retirados pois diminui assim a possibilidade de redundância e para facilitar a leitura do comentário pelo algoritmo na faze de stemização. Stop words formam uma lista de termos não representativos para um documento, geralmente essa lista é composta por: preposições, artigos, advérbios, números, pronomes e pontuação. No desenvolvimento desta pesquisa foi utilizada a relação de Stopwords para o português do Brasil propostas por Balinski (2002).
  + Tokenização: Depois de feita a limpeza realizou-se a tokenização, que é utilizada para decompor o documento em cada termo que o compõe. Os delimitadores utilizados para tokenização foram os espaços em branco entre os termos, alocando, assim, cada termo em uma posição do array correspondente ao seu corpus.
  + Stemização: Nesta etapa foi construído em linguagem C um software complementar para realizar a stemização usando o algoritmo RSLP desenvolvido por Orengo (2001), tal algoritmo tem por finalidade a radicalização de determinado termo para reduzir as redundâncias que existiria no corpus com a presença de vários termos flexionados, tornando, assim, o cálculo da frequência ineficiente.
* Contagem dos termos: Estando assim os termos restantes radicalizados foi realizada a contagem dos mesmos, calculando o número de ocorrências de cada termo, deixando, assim, em uma posição de um vetor o termo e seu número de ocorrências como mostrado na figura 10.

Figura 10 - Contagem dos termos.

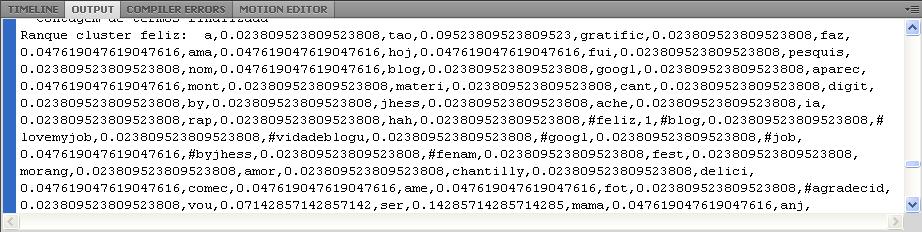


Fonte: Elaborado pelo autor.

## 7.2 FORMAÇÃO DE UM RANQUE DE POLARIDADE DE EXPRESSÕES

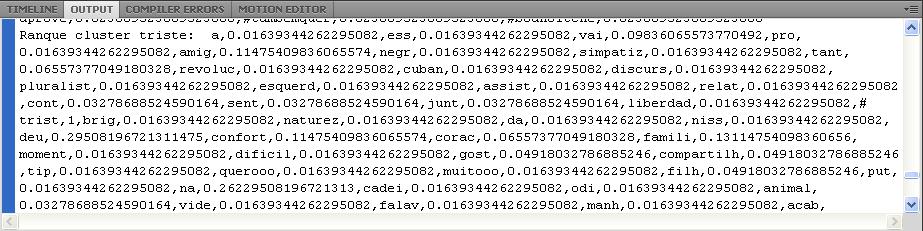
Foi formado um ranque de polaridade conforme se observa nas figuras 11 e 12 abaixo o ranque dos clusters feliz e triste respectivamente, indicando as palavras ou expressões que tem maior importância/influência dentro dos clusters. Cada termo recebeu um peso proporcional à sua frequência. O termo de maior frequência, no cluster positivo, recebeu peso 1 e os demais pesos proporcionais. No caso do cluster negativo o “maior” peso será -1. A escolha de pesos entre 0 e 1 (positivos e negativos) é comumente usada em tarefas de atribuição de importância de termos em pesquisas de processamento de língua natural, justificando, portanto, sua utilização nesta pesquisa. O cluster feliz contem exatamente 283 termos tratados, e o triste 525 como a figura 13 ilustra. A lista dos termos dos clusters feliz e triste e seus respectivos pesos se encontram em anexo a este trabalho, nos anexos B e C respectivamente.

Figura 11 - Ranque de pesos dos termos do Cluster feliz.



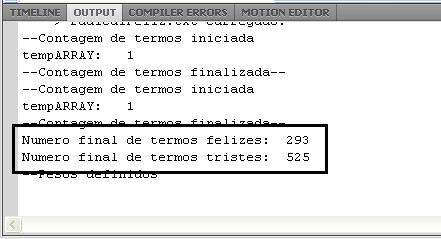
Fonte: Elaborado pelo autor.

Figura 12 - Ranque de pesos dos termos do Cluster triste.



Fonte: Elaborado pelo autor.

Figura 13 - Quantidade de termos nos clusters feliz e triste.



Fonte: Elaborado pelo autor.

## 7.3 CLASSIFICAÇÃO DE COMENTÁRIOS

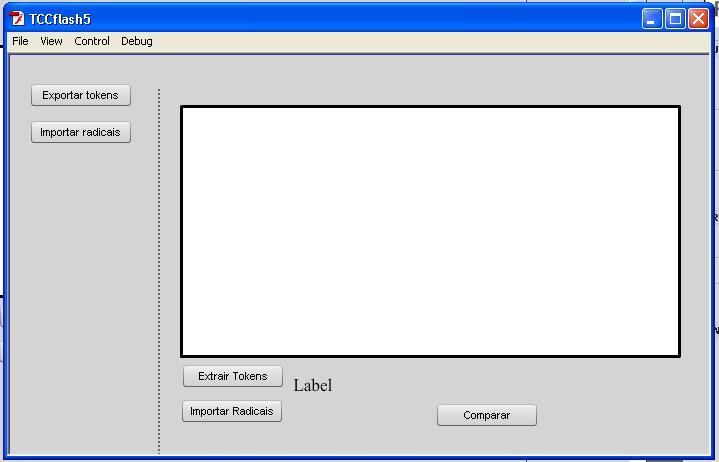
Para a classificação de comentários que não estão dentro do corpus, foi realizado o mesmo tratamento descrito anteriormente: de leitura, extração e limpeza, e contagem dos termos para, assim, poder realizar uma comparação de termos do novo comentário com os termos existentes nos corpus, contando como positivo e/ou negativo a soma final com os pesos proporcionais à frequência de tal termo para determinar, por fim, qual a classificação que o termo receberia, menor que 0 para comentários tristes, maior que 0 para comentários felizes e igual a 0 para comentários que não foi possível a classificação. Esta etapa baseou-se na abordagem na qual um determinado conteúdo é avaliado, considerando o resultado da avaliação de suas palavras, tendo como base um conjunto de palavras previamente classificadas, conforme descrito anteriormente. Nesta técnica existe um dicionário de palavras positivas e negativas, por exemplo: “bom”, “legal”, “ótimo” são palavras positivas, já “péssimo”, “ruim”, “horrível”, são palavras que expressam um sentimento negativo. A vantagem desta técnica está na sua simplicidade, e o fato de não ser necessário classificar previamente um conjunto de conteúdo para treinamento. No entanto, o conjunto de palavras deve ser cuidadosamente selecionado, tratando ainda a ambiguidade de uma mesma palavra em um mesmo tópico. Por exemplo, a palavra ”inveja”, em certo contexto significa uma “inveja boa”, e em outro contexto “inveja ruim”, podendo representar tanto um sentimento positivo, quanto um sentimento negativo. Como forma de verificar a validade do método proposto, foi desenvolvido um protótipo de um sistema de classificação automática de sentimentos com a linguem de programação Action Script 3.0 e com um banco de dados em arquivo XML, com um software aplicativo auxiliar em linguagem C para a realização da técnica de Stemming. O sistema foi testado, verificando o seu índice de acertos (precisão) quanto à polaridade de comentários extraídos de redes sociais. Para tanto, a classificação do sistema para determinado comentário é confrontada com a classificação manual feita por juízes humanos. Os resultados obtidos são descritos na seção seguinte.

**8 RESULTADOS**

Para obtenção dos resultados foram realizados vários testes de forma unitária com uma coleção de 40 novos comentários assim divididos: 10 comentários tristes do Facebook e 10 felizes, da mesma forma com o Twitter (10 tristes e 10 felizes). Cabe destacar que todos os comentários foram coletados manualmente nas duas redes sociais analisadas. Os testes foram realizados em um software desenvolvido para este trabalho como se pode ver nas Figuras 14, 15, 16, e 17. Os comentários utilizados continham #’s (Hashtags) vinculadas com palavras de sentimentos tristes (negativos) e com sentimentos felizes (positivos). Para efeito de comparação e teste da potencialidade da técnica utilizada, cada comentário foi testado duas vezes, uma vez com a # (Hashtag) indicativa e outra sem o referido indicativo. Além dos testes, visando o cálculo da precisão, os dados obtidos foram separados por origem e sentimento, fazendo um comparativo completo do reconhecimento.

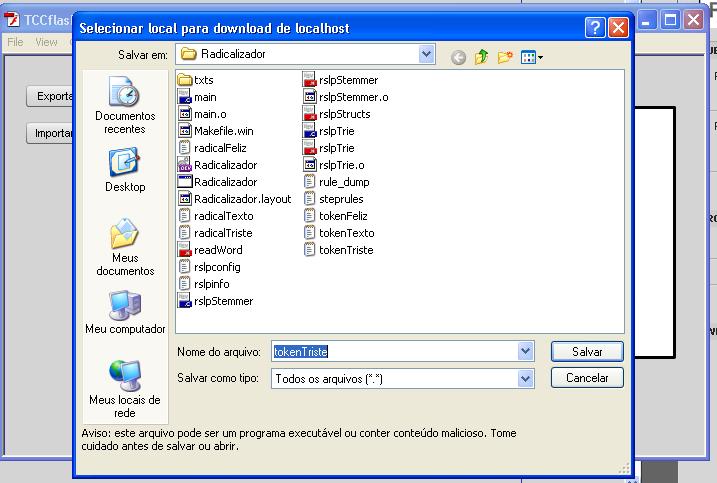
O funcionamento do software construído demanda um pouco de trabalho manual, pois sua arquitetura inicial foi pensada para a finalidade de consulta de um único post por vez, como se observa na figura 14, onde se observa os botões de exportar tokens para a stemização e importar os radicais depois de passados pelo devido tratamento em uma aplicação construída com base no algoritmo para Stemitizar os termos como podemos ver a exportação nas figuras 15, depois de importar os termos já radicalizados é feito o mesmo tratamento manual para o comentário a ser analisado, com os respectivos botões de exportar tokens, figura 16, e importar radicais e, por fim, o botão de comparar que faz uma comparação dos termos recuperados e tratados do comentário a ser analisado com os termos importados e tratados com o corpus como a figura 17 mostra.

Figura 14 - Software de reconhecimento de sentimentos



Fonte: Elaborado pelo autor.

Figura 15 - Software de reconhecimento de sentimentos exportando arquivo tokensTriste.txt.



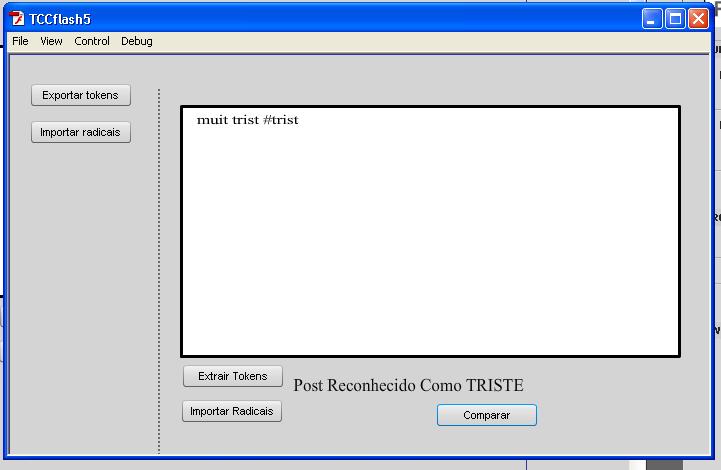
Fonte: Elaborado pelo autor.

Figura 16 - Software de reconhecimento de sentimentos exportando arquivo tokenTexto.txt



Fonte: Elaborado pelo autor.

Figura 17 - Software de reconhecimento de sentimentos reconhecendo o sentimento expresso em determinado comentário já devidamente tratado.



Fonte: Elaborado pelo autor.

Como se pode ver nas figuras 18 e 19 representadas, foi obtido um resultado de 100% de acerto nos testes com o uso de #(Hashtag) indicativa de tristeza e 90% de acerto nos testes sem o uso dessas #s(Hashtags), acredita-se que esta margem de erro ocorra pelo tamanho reduzido do corpus em geral, pois quão maior o tamanho do corpus mais possibilidades de reconhecimento haverá, e também pelo tamanho dos comentários analisados, pois com base nas experimentações feitas foi observado de forma empírica que quanto mais termos o comentário possuir maior vai ser a probabilidade de erro mas, por outro lado, quanto menos termos possuir mais acertada será a análise.

Figura 18 - Gráfico de resultados do reconhecimento de comentários do Facebook com #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Figura 19 - Gráfico de resultados do reconhecimento de comentários do Facebook sem #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Os testes com o Twitter para comentários considerados tristes foram tão satisfatórios quanto os testes para o Facebook, como observados nas Figuras 20 e 21. Obteve-se 100% de acerto tanto para os comentários com #s(Hashtags) como para os sem este indicativo.

Figura 20 - Gráfico de resultados do reconhecimento de comentários do Twitter com #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Figura 21 - Gráfico de resultados do reconhecimento de comentários do Twitter sem #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Analisando todos os dados obtidos nos testes para comentários tristes originados do Facebook e Twitter pode-se perceber uma taxa de acerto alta conforme anteriormente apresentados nas figuras 18, 19, 20, 21. Assim, foi realizada uma nova análise, considerando a média simples de acertos com e sem a utilização de #(Hashtag) indicativa de sentimento triste (#triste), conforme apresentado na figura 22 nota-se que a técnica tem 97% de precisão para os sentimentos tristes.

Figura 22 - Gráfico da média simples dos resultados do reconhecimento de comentários tristes do Twitter e Facebook com e sem #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Pode-se fazer um comparativo nas figuras 23 e 24 onde se obteve um resultado não tão satisfatório com os comentários obtidos no Facebook que expressavam sentimentos de felicidade, pois nos testes onde a #(Hashtag) indicativa de sentimento foi omitida houve uma taxa de erro de 60%, muito acima do esperado. Por outro lado, os testes em que foram mantidas as #s(Hashtags) indicativas de sentimentos, obtiveram 100% de acerto no cálculo de sentimentos. Neste caso acredita-se que essa taxa de erro grande se deva ao tamanho reduzido em especial do cluster feliz, que contem 293 termos finais tratados, em compensação dos 525 termos finais e tratados do cluster triste.

Figura 23 - Gráfico de resultados obtidos dos comentários felizes do Facebook com #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Figura 24 - Gráfico de resultados do reconhecimento de comentários felizes recuperados no Facebook sem #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

O mesmo comportamento observado nos testes com comentários felizes obtidos no Facebook se repetiu com o Twitter, conforme se constata nas figuras 25 e 26. Nos testes em que foram mantidas as #s(Hashtags) indicativas de sentimento, a precisão foi 100% de reconhecimento, mas os testes onde estes indicativos foram retirados houve uma queda considerável do rendimento para 60% de comentários reconhecidos devido ao numero infelior de termos do cluster feliz. Entretanto quando se faz uma média simples dos dados obtidos com esses testes tem-se ainda um resultado satisfatório de 75% de reconhecimento, como se observa na figura 27.

Figura 25 - Gráfico de resultados obtidos no reconhecimento de comentários do Twitter com #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Figura 26 - Gráfico de resultados obtidos no reconhecimento de comentários do Twitter sem #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Figura 27 - Gráfico da média simples dos resultados do reconhecimento de comentários de sentimento de felicidade do Twitter e Facebook com e sem #(Hashtag) indicativa de sentimento.

Fonte: Elaborado pelo autor.

Figura 28 - Gráfico dos resultados do reconhecimento dos comentários tristes e felizes com e sem #(Hashtag) do Facebook e do Twitter.

Fonte: Elaborado pelo autor.

E, por fim, a figura 20 mostra a média total de todos os teste realizados demonstrando que o software reconheceu 86% dos comentários, considerando as diferentes configurações. Sabe-se que devido ao corpus de treinamento e do corpus de teste serem pequenos pode ocorrer variações na precisão alcançada, mas com base neste estudo acredita-se que a técnica aplicada apresenta um bom potencial.

Figura 29 - Gráfico dos resultados do reconhecimento de todos os comentários com todas as possibilidades de teste.

Fonte: Elaborado pelo autor.

# 9 CONSIDERAÇÕES FINAIS

Considerando os resultados obtidos a partir da investigação aqui relatada, pode-se destacar a sua contribuição para pesquisas no campo do processamento de língua natural com foco em língua portuguesa, uma vez que o método proposto obteve uma taxa de precisão de 86% na classificação dos comentários submetidos ao sistema, o que pode ser indicar o potencial da abordagem proposta. Cabe destacar também que tais resultados tornam-se ainda mais interessantes se for considerada a simplicidade da técnica utilizada.

Porém o presente método também apresentou limitações, como as indicadas durante a discussão dos resultados, por exemplo, a que se refere ao processo de coleta do repositório de comentários para formação do corpus. Sabe-se que devido ao corpus de treinamento e do corpus de teste serem pequenos pode haver alguma oscilação nesta apuração para cima ou para baixo, mas com base neste estudo. Como trabalhos futuros com a mesma temática sugere-se a utilização de outro repositório, mais consistente, com número substancialmente maior de opiniões. Também poderá ser realizada por especialistas uma coleta mais seletiva das opiniões utilizadas para o corpus, que pode melhorar a qualidade do repositório.

Outra limitação encontrada refere-se ao processo de desenvolvimento do software que faz uso método proposto. Há uma necessidade de aperfeiçoamento posterior deste software como trabalho futuro, deixando-o mais automatizado e funcional para testes em grande escala. Poder-se-ia, por exemplo, automatizar o processo de coleta do corpus de treinamento nas referidas redes sociais. Outra possibilidade é fazer uma comparação do treinamento realizado neste trabalho com um treinamento realizado com outras teorias, a exemplo de uma rede neural que se enquadre nas necessidades do projeto.

Deste modo, conclui-se que o método proposto e implementado por meio de um software, apesar de simples, tem um grande potencial, já que foram alcançados bons índices de precisão ao se analisar o sentimento associados aos comentários, sendo pouco custoso já que trabalha apenas no nível superficial do texto. Deste modo, o presente trabalho contribui para pesquisas ligadas ao processamento de linguagem natural, principalmente em língua portuguesa, uma vez que há poucos trabalhos específicos nesta língua. Porém ainda há muito a ser melhorado e novas propostas podem contribuir ainda mais para o crescimento de pesquisas na área.

# REFERÊNCIAS

AMORIN, T. Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados. **Centro de Informática UFPE.** Disponível em: <http://www2.cin.ufpe.br/site/index.php>. Acesso em: 19 de Nov de 2014

BALINSKI, R. Filtragem de Informações no Ambiente do Direto. **Instituto de Informática UFRGS**. Disponível em: <http://www.inf.ufrgs.br/procpar/direto/trabalhos/dissertacao-pdf.PDF>. Acesso em: 16 de Maio de 2005.

FACEBOOK. **Company Info | Facebook Newsroom**, Disponível em: <http://newsroom.fb.com/company-info/>. Acesso em: 24 mar. 2014.

GONZALEZ, M; Lima, V. L. S. Recuperação de Informação e Processamento da Linguagem Natural, **PUCRS | Faculdade de Informática, c2007-2008.** Disponível em: <http://www.inf.pucrs.br/~gonzalez/docs/minicurso-jaia2003.pdf>. Acesso em: 25 jun 2015.

LUGER, G. F. **Inteligência artificial: estruturas e estratégias para a solução de problemas complexos.** 5. ed. Porto Alegre: Bookman, 2005.

**MCCARTHY, J. et al.** A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. **Stanford University**, 1955. Disponível em: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>. Acesso em: 25 jun 2015.

ORENGO, V, M; COELHO, A, R; BURIOL, L, S. Removedor de Sufixos da Língua Portuguesa – RSLP, **Institudo de Informática UFRGS,** 2001 Disponível em: <http://www.inf.ufrgs.br/~arcoelho/rslp/html/>. Acesso em: 15 Nov 2014.

O'DELL, J. **How Big Is the Web & How Fast Is It Growing?**, Disponível em: <https://mashable.com/2011/06/19/how-many-websites/>. Acesso em: 5 mar. 2013.

PANG, B. ; LEE, L. Opinion mining and sentiment analysis. **Cornell University Department of Computer Science, 2008**. Disponivel em: <http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf>. Acesso em: 25 jun 2014

PRESSMAN, R. S. **Engenharia de Software: Uma abordagem profissional**. 7ªe. Porto Alegre: AMGH Editora Ltda, 2011.

RODRUIES, C.A.S et al. Mineração de Opinião / Análise de Sentimentos. **Departamento de Informática e Estatística UFSC**, c1960-2010. Disponível em: <http://www.inf.ufsc.br/~alvares/INE5644/MineracaoOpiniao.pdf>. Acesso em 22 ago 2014

SARDINHA, T. B. **Lingüística de corpus**. São Paulo: Manole, 2005.

SILVA, B. C. et al. **Introdução ao processamento das línguas naturais e algumas aplicações.** 2007. 121 f. (Série de Relatórios Técnicos do NILC), São Carlos, 2007. Disponível em: <http://www.letras.etc.br/ebralc/NILCTR0710-DiasDaSilvaEtAl.pdf>. Acesso em: 01 maio 2015.

TWITTER. **Sobre o Twitter, 2014.** Apresenta informações estatísticas sobre a rede social. Disponível em: <https://about.Twitter.com/pt/company>. Acesso em: 24 maio 2015.

WINOGRAD, T. **Understanding natural language**. New York: Academic Press, 1972.

# ANEXO A

Lista de Stopwords utilizadas neste trabalho:

A, a, à, agora, ainda, alguém, algum, alguma, algumas, alguns, ampla, amplas, amplo, amplos, ante, antes, ao, aos, após, aquela, aquelas, aquele, aqueles, aquilo, as, até, ate, através, cada, coisa, coisas, com, como, contra, contudo, da, daquele, daqueles, das, de, dela, delas, dele, deles, depois, dessa, dessas, desse, desses, desta, destas, deste, deste, destes, deve, devem, devendo, dever, deverá, deverão, deveria, deveriam, devia, deviam, disse, disso, disto, dito, diz, dizem, do, dos, E, e, é, ela, elas, ele, eles, em, enquanto, entre, era, essa, essas, esse, esses, esta, está, estamos, estão, estao, estas, estava, estavam, estávamos, estavamos, este, estes, estou, eu, fazendo, fazer, feita, feitas, feito, feitos, foi, foram, fosse, fossem, rande, randes, há, ha, já, ja, la, la, lá, lhe, lhes, lo, mas, me, mesma, mesmas, mesmo, mesmos, meu, meus, minha, minhas, muita, muitas, muito, nao, não, nas, nem, nenhum, nessa, nessas, nesta, nestas, ninguem, no, nos, nós, nossa, nossas, nosso, nossos ,num, numa, nunca, o, os, outra, outras, outro, outros, para, pela, pelas, pelo, pelos, pequena, pequenas, pequeno, pequenos, per, perante, pode, pôde, podendo, poder, poderia, poderiam, podia, podiam, pois, por, porém, porque , posso, pouca, poucas, pouco, poucos, primeiro, primeiros, própria, propria, próprias, próprias, próprio, proprio, próprios, proprios, quais, qual, quando, quanto, quantos, que, quem, são, são, se, seja, sejam, sem, sempre, sendo, será, serão, seu, seus, si, sido, só, sob, sobre, sua, suas, talvez, também, tampouco, te, tem, tendo, tenha, ter, teu, teus, ti, tido, tinha, tinham, toda, todas, todavia, todo, todos, tu, tua, tuas, tudo, última, ultima, últimas, ultimas, último, ultimo, últimos, ultimos, um, uma, umas, uns, vendo, ver, vez, vindo, vir, vos, vós, isso, isto.

# ANEXO B

Lista de termos já tratados e separados com seus respectivos termos e pesos, organizados em ordem crescente de peso:

#feliz,1

vou,0.07142857142857142

ceu,0.07142857142857142

bem,0.07142857142857142

amig,0.07142857142857142

quer,0.07142857142857142

tenh,0.07142857142857142

feliz,0.07142857142857142

nad,0.07142857142857142

famili,0.07142857142857142

acredit,0.07142857142857142

tao,0.09523809523809523

boa,0.09523809523809523

noit,0.09523809523809523

jesu,0.09523809523809523

na,0.09523809523809523

bom,0.09523809523809523

to,0.09523809523809523

pra,0.11904761904761904

#am,0.11904761904761904

vai,0.11904761904761904

ser,0.14285714285714285

maol,0.14285714285714285

am,0.14285714285714285

obrig,0.14285714285714285

ano,0.16666666666666666

dia,0.19047619047619047

vid,0.19047619047619047

deu,0.21428571428571427

a,0.023809523809523808

gratific,0.023809523809523808

fui,0.023809523809523808

pesquis,0.023809523809523808

blog,0.023809523809523808

googl,0.023809523809523808

mont,0.023809523809523808

materi,0.023809523809523808

cant,0.023809523809523808

digit,0.023809523809523808

by,0.023809523809523808

jhess,0.023809523809523808

ache,0.023809523809523808

ia,0.023809523809523808

rap,0.023809523809523808

hah,0.023809523809523808

#blog,0.023809523809523808

#lovemyjob,0.023809523809523808

#vidadeblogu,0.023809523809523808

#googl,0.023809523809523808

#byjhess,0.023809523809523808

#fenam,0.023809523809523808

fest,0.023809523809523808

morang,0.023809523809523808

amor,0.023809523809523808

chantilly,0.023809523809523808

fot,0.023809523809523808

#agradecid,0.023809523809523808

troux,0.023809523809523808

bonit,0.023809523809523808

joi,0.023809523809523808

perfeit,0.023809523809523808

cuid,0.023809523809523808

got,0.023809523809523808

cristalin,0.023809523809523808

inocenci,0.023809523809523808

escolh,0.023809523809523808

sere,0.023809523809523808

port,0.023809523809523808

guardi,0.023809523809523808

purez,0.023809523809523808

#sofi,0.023809523809523808

namor,0.023809523809523808

sim,0.023809523809523808

#deusabencoeonasc,0.023809523809523808

vind,0.023809523809523808

fernand,0.023809523809523808

prepar,0.023809523809523808

esforc,0.023809523809523808

correri,0.023809523809523808

permit,0.023809523809523808

estare,0.023809523809523808

di,0.023809523809523808

benc,0.023809523809523808

fiqu,0.023809523809523808

esp,0.023809523809523808

#muit,0.023809523809523808

hyunda,0.023809523809523808

motor,0.023809523809523808

brasil,0.023809523809523808

barrig,0.023809523809523808

pesso,0.023809523809523808

lad,0.023809523809523808

#bomdi,0.023809523809523808

seg,0.023809523809523808

flux,0.023809523809523808

mar,0.023809523809523808

sobrenom,0.023809523809523808

cheir,0.023809523809523808

so,0.023809523809523808

hc,0.023809523809523808

pneu,0.023809523809523808

adore,0.023809523809523808

surpr,0.023809523809523808

#top,0.023809523809523808

bess,0.023809523809523808

corac,0.023809523809523808

complet,0.023809523809523808

parec,0.023809523809523808

fic,0.023809523809523808

lind,0.023809523809523808

kkkkkkk,0.023809523809523808

brinc,0.023809523809523808

diz,0.023809523809523808

descrev,0.023809523809523808

apen,0.023809523809523808

palavr,0.023809523809523808

seri,0.023809523809523808

conced,0.023809523809523808

mudanc,0.023809523809523808

conquist,0.023809523809523808

pen,0.023809523809523808

final,0.023809523809523808

23,0.023809523809523808

aninh,0.023809523809523808

acontec,0.023809523809523808

#parabenspramim,0.023809523809523808

#eulind,0.023809523809523808

#23,0.023809523809523808

perd,0.023809523809523808

chor,0.023809523809523808

atr,0.023809523809523808

fiz,0.023809523809523808

do,0.023809523809523808

sang,0.023809523809523808

#doarsang,0.023809523809523808

sab,0.023809523809523808

solt,0.023809523809523808

sorr,0.023809523809523808

toa,0.023809523809523808

jeit,0.023809523809523808

quol,0.023809523809523808

#solt,0.023809523809523808

#alegr,0.023809523809523808

abrac,0.023809523809523808

forc,0.023809523809523808

pro,0.023809523809523808

niss,0.023809523809523808

louv,0.023809523809523808

#sab,0.023809523809523808

diaa,0.023809523809523808

companhi,0.023809523809523808

test,0.023809523809523808

present,0.023809523809523808

#ame,0.023809523809523808

#pres,0.023809523809523808

#fotobast,0.023809523809523808

#inlovefot,0.023809523809523808

dupl,0.023809523809523808

dinam,0.023809523809523808

robert,0.023809523809523808

#amomeutrabalh,0.023809523809523808

#mus,0.023809523809523808

#sax,0.023809523809523808

aproveit,0.023809523809523808

relax,0.023809523809523808

import,0.023809523809523808

#herbalif,0.023809523809523808

est,0.023809523809523808

motiv,0.023809523809523808

melhor,0.023809523809523808

sonh,0.023809523809523808

#mex,0.023809523809523808

#cancun,0.023809523809523808

#xcaret,0.023809523809523808

tortuguita,0.023809523809523808

#chocolat,0.023809523809523808

#tortuguit,0.023809523809523808

sou,0.023809523809523808

mund,0.023809523809523808

dar,0.023809523809523808

precis,0.023809523809523808

encontr,0.023809523809523808

amanh,0.023809523809523808

#mar,0.023809523809523808

vem,0.023809523809523808

saudad,0.023809523809523808

ta,0.023809523809523808

enorm,0.023809523809523808

tant,0.023809523809523808

convers,0.023809523809523808

falt,0.023809523809523808

temp,0.023809523809523808

sent,0.023809523809523808

encerr,0.023809523809523808

#complet,0.023809523809523808

#como,0.023809523809523808

#omeuam,0.023809523809523808

#porvoc,0.023809523809523808

#eu,0.023809523809523808

#simpl,0.023809523809523808

#meam,0.023809523809523808

sal,0.023809523809523808

conclu,0.023809523809523808

sucess,0.023809523809523808

#pilat,0.023809523809523808

#fisioterapi,0.023809523809523808

fat,0.023809523809523808

jamal,0.023809523809523808

pod,0.023809523809523808

ignor,0.023809523809523808

otim,0.023809523809523808

seman,0.023809523809523808

vc,0.023809523809523808

bj,0.023809523809523808

#look,0.023809523809523808

#lux,0.023809523809523808

#lind,0.023809523809523808

#moren,0.023809523809523808

#fashion,0.023809523809523808

#dem,0.023809523809523808

#ador,0.023809523809523808

#fest,0.023809523809523808

reluz,0.023809523809523808

cabel,0.023809523809523808

art,0.023809523809523808

alem,0.023809523809523808

#happy,0.023809523809523808

#umdiaqualqu,0.023809523809523808

#saturday,0.023809523809523808

#sabadin,0.023809523809523808

ont,0.023809523809523808

#amig,0.023809523809523808

#liind,0.023809523809523808

#vivendosemexplicac,0.023809523809523808

#vamoquev,0.023809523809523808

suced,0.023809523809523808

#sucess,0.023809523809523808

#t,0.023809523809523808

gost,0.023809523809523808

#churrasc,0.023809523809523808

#praiaclubesaofrancisc,0.023809523809523808

#nitero,0.023809523809523808

#comemorac,0.023809523809523808

#paz,0.023809523809523808

#memori,0.023809523809523808

prim,0.023809523809523808

#miss,0.023809523809523808

#adamantin,0.023809523809523808

#sacerdoci,0.023809523809523808

#cancaonov,0.023809523809523808

ach,0.023809523809523808

sorris,0.023809523809523808

#carpedi,0.023809523809523808

vera,0.023809523809523808

cheg,0.023809523809523808

#peacelif,0.023809523809523808

#prai,0.023809523809523808

#aprove,0.023809523809523808

#tambemquer,0.023809523809523808

#boanoitche,0.023809523809523808

#boanoitche,0.023809523809523808

#212club,0.023809523809523808

faz,0.047619047619047616

ama,0.047619047619047616

hoj,0.047619047619047616

nom,0.047619047619047616

aparec,0.047619047619047616

#job,0.047619047619047616

delici,0.047619047619047616

comec,0.047619047619047616

ame,0.047619047619047616

mama,0.047619047619047616

anj,0.047619047619047616

mim,0.047619047619047616

felic,0.047619047619047616

doming,0.047619047619047616

ped,0.047619047619047616

cas,0.047619047619047616

abenco,0.047619047619047616

mai,0.047619047619047616

agradec,0.047619047619047616

maravilh,0.047619047619047616

trabalh,0.047619047619047616

pass,0.047619047619047616

part,0.047619047619047616

val,0.047619047619047616

volt,0.047619047619047616

dig,0.047619047619047616

voc,0.047619047619047616

tbm,0.047619047619047616

nov,0.047619047619047616

realiz,0.047619047619047616

crist,0.047619047619047616

assim,0.047619047619047616

# ANEXO C

#trist,1

a,0.01639344262295082

ess,0.01639344262295082

pro,0.01639344262295082

negr,0.01639344262295082

simpatiz,0.01639344262295082

revoluc,0.01639344262295082

cuban,0.01639344262295082

discurs,0.01639344262295082

pluralist,0.01639344262295082

esquerd,0.01639344262295082

assist,0.01639344262295082

relat,0.01639344262295082

liberdad,0.01639344262295082

brig,0.01639344262295082

naturez,0.01639344262295082

da,0.01639344262295082

niss,0.01639344262295082

moment,0.01639344262295082

dificil,0.01639344262295082

tip,0.01639344262295082

querooo,0.01639344262295082

muitooo,0.01639344262295082

put,0.01639344262295082

cadei,0.01639344262295082

odi,0.01639344262295082

vide,0.01639344262295082

falav,0.01639344262295082

manh,0.01639344262295082

pres,0.01639344262295082

#horrorizada#,0.01639344262295082

bia,0.01639344262295082

espec,0.01639344262295082

es,0.01639344262295082

nils,0.01639344262295082

tristez,0.01639344262295082

valeri,0.01639344262295082

crianc,0.01639344262295082

alun,0.01639344262295082

doent,0.01639344262295082

uti,0.01639344262295082

tev,0.01639344262295082

vari,0.01639344262295082

convulco,0.01639344262295082

sext,0.01639344262295082

feir,0.01639344262295082

ajud,0.01639344262295082

orac,0.01639344262295082

arras,0.01639344262295082

#choc,0.01639344262295082

#preocup,0.01639344262295082

form,0.01639344262295082

agir,0.01639344262295082

louv,0.01639344262295082

or,0.01639344262295082

testemunh,0.01639344262295082

serv,0.01639344262295082

anim,0.01639344262295082

estimul,0.01639344262295082

bel,0.01639344262295082

muitoo,0.01639344262295082

falec,0.01639344262295082

prim,0.01639344262295082

coraco,0.01639344262295082

#arras,0.01639344262295082

vao,0.01639344262295082

#decoracaopart,0.01639344262295082

revolt,0.01639344262295082

ond,0.01639344262295082

vam,0.01639344262295082

par,0.01639344262295082

instal,0.01639344262295082

dentr,0.01639344262295082

lar,0.01639344262295082

inimig,0.01639344262295082

usa,0.01639344262295082

misericordi,0.01639344262295082

seri,0.01639344262295082

ai,0.01639344262295082

violent,0.01639344262295082

mentir,0.01639344262295082

log,0.01639344262295082

jus,0.01639344262295082

celest,0.01639344262295082

rai,0.01639344262295082

nuv,0.01639344262295082

receber,0.01639344262295082

conform,0.01639344262295082

fez,0.01639344262295082

miste,0.01639344262295082

vej,0.01639344262295082

jeit,0.01639344262295082

and,0.01639344262295082

gerac,0.01639344262295082

misercordi,0.01639344262295082

sensac,0.01639344262295082

batalh,0.01639344262295082

roub,0.01639344262295082

celul,0.01639344262295082

emoc,0.01639344262295082

lament,0.01639344262295082

precoc,0.01639344262295082

anderson,0.01639344262295082

vasconcel,0.01639344262295082

ex,0.01639344262295082

prof,0.01639344262295082

tenol,0.01639344262295082

mes,0.01639344262295082

korpu,0.01639344262295082

boa,0.01639344262295082

esper,0.01639344262295082

georg,0.01639344262295082

irma,0.01639344262295082

venh,0.01639344262295082

ora,0.01639344262295082

clam,0.01639344262295082

princip,0.01639344262295082

amo,0.01639344262295082

prest,0.01639344262295082

cirurgi,0.01639344262295082

alt,0.01639344262295082

risc,0.01639344262295082

desej,0.01639344262295082

tenh,0.01639344262295082

malign,0.01639344262295082

dig,0.01639344262295082

autoridad,0.01639344262295082

conced,0.01639344262295082

determin,0.01639344262295082

cur,0.01639344262295082

nest,0.01639344262295082

crei,0.01639344262295082

far,0.01639344262295082

for,0.01639344262295082

honr,0.01639344262295082

glori,0.01639344262295082

#muitotr,0.01639344262295082

control,0.01639344262295082

hom,0.01639344262295082

assum,0.01639344262295082

respons,0.01639344262295082

ato,0.01639344262295082

transfer,0.01639344262295082

culp,0.01639344262295082

terc,0.01639344262295082

caminh,0.01639344262295082

curt,0.01639344262295082

incompetenci,0.01639344262295082

supere,0.01639344262295082

imaginav,0.01639344262295082

capaz,0.01639344262295082

realiz,0.01639344262295082

sucess,0.01639344262295082

fiz,0.01639344262295082

chat,0.01639344262295082

receberi,0.01639344262295082

bol,0.01639344262295082

frent,0.01639344262295082

dras,0.01639344262295082

facil,0.01639344262295082

julg,0.01639344262295082

jov,0.01639344262295082

morfin,0.01639344262295082

faz,0.01639344262295082

efeit,0.01639344262295082

mat,0.01639344262295082

doi,0.01639344262295082

oportun,0.01639344262295082

fat,0.01639344262295082

complet,0.01639344262295082

decepcion,0.01639344262295082

contra,0.01639344262295082

repar,0.01639344262295082

vez,0.01639344262295082

surpreend,0.01639344262295082

explicac,0.01639344262295082

sentire,0.01639344262295082

raf,0.01639344262295082

veloc,0.01639344262295082

son,0.01639344262295082

combin,0.01639344262295082

conex,0.01639344262295082

report,0.01639344262295082

motorist,0.01639344262295082

motiv,0.01639344262295082

mot,0.01639344262295082

fortalez,0.01639344262295082

ces,0.01639344262295082

trabalh,0.01639344262295082

assemblei,0.01639344262295082

legisl,0.01639344262295082

esp,0.01639344262295082

trabalhav,0.01639344262295082

region,0.01639344262295082

assemble,0.01639344262295082

cruz,0.01639344262295082

dumont,0.01639344262295082

via,0.01639344262295082

express,0.01639344262295082

amanhec,0.01639344262295082

cinz,0.01639344262295082

dolor,0.01639344262295082

bem,0.01639344262295082

acord,0.01639344262295082

pass,0.01639344262295082

mal,0.01639344262295082

#agoni,0.01639344262295082

#coracaoapert,0.01639344262295082

math,0.01639344262295082

andrion,0.01639344262295082

parc,0.01639344262295082

consum,0.01639344262295082

relacion,0.01639344262295082

presenci,0.01639344262295082

lastima,0.01639344262295082

metr,0.01639344262295082

transex,0.01639344262295082

vaga,0.01639344262295082

feminin,0.01639344262295082

grup,0.01639344262295082

mulh,0.01639344262295082

ria,0.01639344262295082

descontrol,0.01639344262295082

us,0.01639344262295082

cuec,0.01639344262295082

enfi,0.01639344262295082

preferenci,0.01639344262295082

fal,0.01639344262295082

sou,0.01639344262295082

obrig,0.01639344262295082

respit,0.01639344262295082

sim,0.01639344262295082

vagn,0.01639344262295082

souz,0.01639344262295082

leonel,0.01639344262295082

cresc,0.01639344262295082

pow,0.01639344262295082

consegu,0.01639344262295082

cheg,0.01639344262295082

local,0.01639344262295082

dev,0.01639344262295082

tempor,0.01639344262295082

cert,0.01639344262295082

trat,0.01639344262295082

brinqued,0.01639344262295082

algu,0.01639344262295082

just,0.01639344262295082

derram,0.01639344262295082

pitad,0.01639344262295082

terr,0.01639344262295082

uma,0.01639344262295082

gat,0.01639344262295082

docil,0.01639344262295082

brinc,0.01639344262295082

sincer,0.01639344262295082

monstr,0.01639344262295082

queri,0.01639344262295082

surt,0.01639344262295082

fof,0.01639344262295082

cal,0.01639344262295082

ahhhh,0.01639344262295082

feliz,0.01639344262295082

sabi,0.01639344262295082

adiant,0.01639344262295082

cerebr,0.01639344262295082

torn,0.01639344262295082

depend,0.01639344262295082

tel,0.01639344262295082

tv,0.01639344262295082

ach,0.01639344262295082

maconh,0.01639344262295082

coment,0.01639344262295082

mand,0.01639344262295082

chei,0.01639344262295082

estud,0.01639344262295082

ter,0.01639344262295082

band,0.01639344262295082

rbd,0.01639344262295082

lollapalooz,0.01639344262295082

car,0.01639344262295082

limit,0.01639344262295082

mens,0.01639344262295082

post,0.01639344262295082

red,0.01639344262295082

soc,0.01639344262295082

#boanoit,0.01639344262295082

cad,0.01639344262295082

futur,0.01639344262295082

marid,0.01639344262295082

abandon,0.01639344262295082

gent,0.01639344262295082

bom,0.01639344262295082

😕,0.01639344262295082

melhor,0.01639344262295082

dj,0.01639344262295082

bh,0.01639344262295082

pirt,0.01639344262295082

monic,0.01639344262295082

cqc,0.01639344262295082

mostr,0.01639344262295082

caxi,0.01639344262295082

ma,0.01639344262295082

tax,0.01639344262295082

mortal,0.01639344262295082

infantol,0.01639344262295082

inevita,0.01639344262295082

mo,0.01639344262295082

papa,0.01639344262295082

noel,0.01639344262295082

ne,0.01639344262295082

impression,0.01639344262295082

espor,0.01639344262295082

coloc,0.01639344262295082

segund,0.01639344262295082

plan,0.01639344262295082

amizad,0.01639344262295082

tim,0.01639344262295082

futebol,0.01639344262295082

joa,0.01639344262295082

unic,0.01639344262295082

amanh,0.01639344262295082

polic,0.01639344262295082

reforc,0.01639344262295082

assalt,0.01639344262295082

nov,0.01639344262295082

meg,0.01639344262295082

sen,0.01639344262295082

ray,0.01639344262295082

lembr,0.01639344262295082

esquec,0.01639344262295082

marc,0.01639344262295082

belez,0.01639344262295082

import,0.01639344262295082

dam,0.01639344262295082

desrespeit,0.01639344262295082

excel,0.01639344262295082

profiss,0.01639344262295082

doming,0.01639344262295082

chuv,0.01639344262295082

tent,0.01639344262295082

pov,0.01639344262295082

qr,0.01639344262295082

corrupc,0.01639344262295082

livr,0.01639344262295082

cub,0.01639344262295082

lut,0.01639344262295082

estelionat,0.01639344262295082

eleitor,0.01639344262295082

consequenci,0.01639344262295082

graviss,0.01639344262295082

necessit,0.01639344262295082

mao,0.01639344262295082

#paietern,0.01639344262295082

#jesu,0.01639344262295082

permanec,0.01639344262295082

#mago,0.01639344262295082

#sent,0.01639344262295082

aqu,0.01639344262295082

corr,0.01639344262295082

fech,0.01639344262295082

janel,0.01639344262295082

pergunte,0.01639344262295082

blind,0.01639344262295082

hah,0.01639344262295082

#thisisriodejan,0.01639344262295082

lembranc,0.01639344262295082

convers,0.01639344262295082

apag,0.01639344262295082

fam,0.01639344262295082

unfollow,0.01639344262295082

quim,0.01639344262295082

escol,0.01639344262295082

acert,0.01639344262295082

questo,0.01639344262295082

part,0.01639344262295082

ano,0.01639344262295082

piad,0.01639344262295082

#aprendinoen,0.01639344262295082

comec,0.01639344262295082

diet,0.01639344262295082

urgent,0.01639344262295082

intensific,0.01639344262295082

exercici,0.01639344262295082

fisic,0.01639344262295082

afff,0.01639344262295082

#drog,0.01639344262295082

#sofaltavaess,0.01639344262295082

#odeiodiet,0.01639344262295082

#adeusdoc,0.01639344262295082

#adeusdoc,0.01639344262295082

11,0.01639344262295082

40,0.01639344262295082

cont,0.03278688524590164

sent,0.03278688524590164

junt,0.03278688524590164

animal,0.03278688524590164

acab,0.03278688524590164

noit,0.03278688524590164

forc,0.03278688524590164

guilherm,0.03278688524590164

desd,0.03278688524590164

pec,0.03278688524590164

explic,0.03278688524590164

dor,0.03278688524590164

jo,0.03278688524590164

exempl,0.03278688524590164

fe,0.03278688524590164

diz,0.03278688524590164

poss,0.03278688524590164

viv,0.03278688524590164

descans,0.03278688524590164

violenci,0.03278688524590164

fort,0.03278688524590164

imagin,0.03278688524590164

som,0.03278688524590164

alg,0.03278688524590164

mort,0.03278688524590164

infeliz,0.03278688524590164

menos,0.03278688524590164

mim,0.03278688524590164

leit,0.03278688524590164

enta,0.03278688524590164

ou,0.03278688524590164

assim,0.03278688524590164

haj,0.03278688524590164

corp,0.03278688524590164

vontad,0.03278688524590164

corag,0.03278688524590164

receb,0.03278688524590164

ced,0.03278688524590164

tard,0.03278688524590164

contin,0.03278688524590164

tom,0.03278688524590164

atitud,0.03278688524590164

vem,0.03278688524590164

conhec,0.03278688524590164

comig,0.03278688524590164

alcool,0.03278688524590164

drog,0.03278688524590164

direc,0.03278688524590164

vit,0.03278688524590164

pastor,0.03278688524590164

paul,0.03278688524590164

temp,0.03278688524590164

deix,0.03278688524590164

assassin,0.03278688524590164

sant,0.03278688524590164

ariel,0.03278688524590164

acontec,0.03278688524590164

cen,0.03278688524590164

cu,0.03278688524590164

brasil,0.03278688524590164

est,0.03278688524590164

human,0.03278688524590164

precis,0.03278688524590164

namor,0.03278688524590164

to,0.03278688524590164

val,0.03278688524590164

morr,0.03278688524590164

cas,0.03278688524590164

bairr,0.03278688524590164

mai,0.03278688524590164

perceb,0.03278688524590164

mud,0.03278688524590164

100,0.03278688524590164

gost,0.04918032786885246

compartilh,0.04918032786885246

filh,0.04918032786885246

ser,0.04918032786885246

quer,0.04918032786885246

acabe,0.04918032786885246

sab,0.04918032786885246

hospit,0.04918032786885246

fique,0.04918032786885246

entend,0.04918032786885246

fic,0.04918032786885246

volt,0.04918032786885246

real,0.04918032786885246

pi,0.04918032786885246

irm,0.04918032786885246

saud,0.04918032786885246

doenc,0.04918032786885246

sai,0.04918032786885246

tao,0.04918032786885246

imprudenci,0.04918032786885246

nad,0.04918032786885246

lamenta,0.04918032786885246

sair,0.04918032786885246

en,0.04918032786885246

hor,0.04918032786885246

tant,0.06557377049180328

corac,0.06557377049180328

paz,0.06557377049180328

mae,0.06557377049180328

notici,0.06557377049180328

histori,0.06557377049180328

prov,0.06557377049180328

am,0.06557377049180328

perd,0.06557377049180328

pai,0.06557377049180328

nom,0.06557377049180328

crist,0.06557377049180328

caus,0.06557377049180328

cidad,0.06557377049180328

pal,0.06557377049180328

voc,0.08196721311475409

pra,0.08196721311475409

sofr,0.08196721311475409

tir,0.08196721311475409

hoj,0.08196721311475409

qu,0.08196721311475409

vai,0.09836065573770492

jesu,0.09836065573770492

vid,0.09836065573770492

mund,0.09836065573770492

senh,0.09836065573770492

saudad,0.09836065573770492

amig,0.11475409836065574

confort,0.11475409836065574

dia,0.11475409836065574

pesso,0.11475409836065574

famili,0.13114754098360656

trist,0.14754098360655737

maol,0.18032786885245902

na,0.26229508196721313

deu,0.29508196721311475

1. Corpus linguístico é um conjunto de textos escritos ou falados numa língua que serve como base de análise. [↑](#footnote-ref-1)
2. Modelo de representação em que os documentos são vistos simplesmente como um “saco de palavras”. [↑](#footnote-ref-2)