UNIVERSIDADE FEDERAL DO PARÁ
CAMPUS UNIVERSITÁRIO DE CASTANHAL
FACULDADE DE SISTEMAS DE INFORMAÇÃO
VANDERSON RUAN PEIXOTO SOARES

ANÁLISE DE SENTIMENTOS AUXILIANDO O PROCESSO DE TOMADA DE DECISÃO: uma abordagem não supervisionada utilizando o recurso léxico *SentiWordNet*

Castanhal – Pará
2018

VANDERSON RUAN PEIXOTO SOARES

ANÁLISE DE SENTIMENTOS AUXILIANDO O PROCESSO DE TOMADA DE DECISÃO: uma abordagem não supervisionada utilizando o recurso léxico *SentiWordNet*

Trabalho de Conclusão de Curso apresentado à Faculdade de Sistemas de Informação do Campus de Castanhal da Universidade Federal do Pará, como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Igor Ruiz Gomes

Castanhal – Pará

2018

VANDERSON RUAN PEIXOTO SOARES

ANÁLISE DE SENTIMENTOS AUXILIANDO O PROCESSO DE TOMADA DE DECISÃO: uma abordagem não supervisionada com o recurso léxico *SentiWordNet*

Trabalho de Conclusão de Curso apresentado à Faculdade de Sistemas de Informação do Campus de Castanhal da Universidade Federal do Pará, como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Igor Ruiz Gomes

BANCA EXAMINADORA

_____

Prof. Dr. Igor Ruiz Gomes (Orientador)

_____

Prof. Dr. Tássio Costa de Carvalho

_____

Prof. Dr. José Jailton Henrique Ferreira Junior

Avaliado em: _____/_____/_____

Conceito:_____

DEDICATÓRIA

Dedico esse trabalho às três mulheres que me inspiram, motivam e me fazem querer ser uma pessoa melhor a cada dia. Minha mãe, Iranilde (Bianca) Soares, por ser minha maior inspiradora de luta, coragem e fé. À minha esposa, Ariana Leal, pela referência, o apoio e a força de uma guerreira e à minha pequena princesa, Sophia Soares, minha filha que é o maior motivo para eu seguir em frente nessa batalha.

AGRADECIMENTOS

O agradecimento é uma forma de reconhecer que as dificuldades da vida não são superadas sozinhas. Cada fase desse caminho só foi possível de ser superada por que tive ao meu lado pessoas maravilhosas.

Assim, agradeço à minha mãe, Iranilde Peixoto Soares (Bianca), que mesmo na simplicidade se esforçou para me dar a melhor educação, sempre me motivando a trilhar bons caminhos.

À minha esposa que esteve comigo durantes estes quatro anos conturbados da faculdade e que sempre foi um ombro amigo, uma conselheira e uma incentivadora.

Agradeço ao meu amigo Rodrigo Ferreira pelas palavras de apoio nos primeiros semestres de faculdade quando, pela primeira vez, pensei em desistir.

Agradeço aos meus amigos mais próximos que estiveram sempre unidos e presentes desde a infância até a fase adulta.

Quero agradecer imensamente aos professores, colabores e colegas do curso, especialmente a turma 2014.2, que de modo geral contribuíram para o meu desenvolvimento acadêmico e profissional.

Ao meu orientador, Prof. Igor, pelo apoio e atenção prestada nas minhas dúvidas e esclarecimentos, nos inúmeros contatos realizados.

Agradeço a todos os membros da banca por dela participarem e pelas contribuições que dela emergirem.

Enfim, agradeço a todos aqueles que estiverem comigo durante estes quatro anos não só em presença, mas também em oração. Obrigado de coração pelo apoio, torcida e orações.

A journey of a thousand miles begins with the first step. And this is not a problem, because now you all know how to walk.

Sid Efromovich – TED

(...), *a man gets what he earns, when he earns it.* (...).

Uncle Benjen – Game of Thrones

Todos nesse país deveriam aprender a programar, pois isso nos ensina a pensar.

Steve Jobs

As coisas que você vê ao seu redor foram feitas por pessoas não mais inteligentes que você, então porque não fazer você mesmo essas coisas.

Steve Jobs

RESUMO

Análise de Sentimentos, também chamada de Mineração de Opinião, explora o estudo computacional de opiniões, sentimentos e emoções expressas em fontes como textos não estruturados. As redes sociais *online* movimentam volumes massivos de dados, hoje muito valorizados por organizações por terem o potencial de gerar informação de significativa relevância para os negócios. A *web* tem se tornado um canal onde os usuários compartilham, explicam ou escrevem sobre suas vidas e interesses, dão opiniões e avaliam a opinião de outros. Essas informações podem vir a influenciar outras pessoas em tomadas de decisão, servindo como base adicional que, frequentemente, não está disponível na descrição. No entanto, essas opiniões estão dispersas na *web*, em formato livre, tornando impossível a busca e análise manual dessas informações, portanto, podemos usar algoritmos para automaticamente classificar milhares de *posts*, sem ter a necessidade de lê-los manualmente. Nesse sentido, esse trabalho visa mostrar que o intuito da inteligência artificial é mostrar para as pessoas que não veio para substituí-las, muito pelo contrário, ela vem para aumentar a inteligência do ser humano e ajudar a processar dados, que hoje, são humanamente impossíveis. Nesse contexto, esse trabalho tem como objetivo analisar a capacidade de predição de um modelo, para análise de sentimentos em textos, desenvolvido para a língua inglesa por Andrea Esuli e Fabrizio Sebastiani: *SentiWordNet*. Esse modelo servirá como base para o desenvolvimento de um sistema para apoio a tomadas de decisão.

Palavras-chave: Análise de sentimentos, mineração de opinião, descoberta de conhecimento, *SentiWordNet*

ABSTRACT

Sentiment Analysis, also called Opinion Mining, explores the computational study of opinions, feelings and emotions expressed in sources as unstructured texts. The online social networks move massive volumes of data, now highly valued by organizations for having the potential to generate information of significant relevance to the business. The web has become a channel where users exchange, explain or write about their lives and interests, give opinions and rate others' opinions. This information may influence others in decision making, serving as an additional basis that is often not available in the description. However, these opinions are scattered on the web, in free format, making it impossible to manually search and analyze this information, hence, we can use algorithms to automatically classify thousands of posts, without having to read them manually. In this sense, this paper aims to show that the aim of artificial intelligence is to show people that it has not come to replace them, quite the contrary, it comes to increase the intelligence of the human being and help to process data, which today are humanly impossible. In this context, this paper aims to analyze the predictive capacity of a model, for sentiment analysis in texts, developed for the English language by Andrea Esuli and Fabrizio Sebastiani: SentiWordNet. This model will serve as the basis for the development of a system to support decision making.

Keywords: Sentiment analysis, opinion mining, knowledge discovery, SentiWordNet

LISTA DE ILUSTRAÇÕES

LISTA DE EQUAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
API	*Application Program Interface*
CSS	*Cascading Style Sheets*
CSV	*Comma Separated Values*
GPL	*General Public License*
HTML	*Hyper Text Markup Language*
IA	Inteligência Artificial
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Database*
KDT	*Knowledge Discovery from Text*
LIWC	*Linguistic Inquiry and Word Count*
MD	Mineração de Dados
MT	Mineração de Textos
MVC	*Model-View-Controller*
NB	*Naive Bayes*
PHP	PHP: *Hypertext Preprocessor*
PLN	Processamento de Linguagem Natural
POS	*Part of Speech*
RI	Recuperação de Informação
RNA	Redes Neurais Artificiais
SBC	Sociedade Brasileira de Computação
SGBD	Sistema de Gerenciamento de Banco de Dados
STF	Superior Tribunal Federal
SVM	*Support Vector Machine*
SWN	*SentiWordNet*
TIC	Tecnologia de Informação e Comunicação
TRF-4	Tribunal Regional Federal da 4ª Região
UFMG	Universidade Federal de Minas Gerais
WWW	*World Wide Web*

SUMÁRIO

1. INTRODUÇÃO

Cada vez mais as Tecnologias de Informação e Comunicação (TICs) têm estimulado a cooperação das ciências, derrubando fronteiras paradigmáticas, institucionais e continentais. “As atividades informatizadas parecem destinadas a melhorar a produtividade nos anos que virão, permitindo assim que continue a crescer a quantidade de informações científicas em circulação” (MEADOWS, 1999, p. 246).

Os avanços tecnológicos na área da informação têm despertado os interesses de diversas áreas do conhecimento na análise de suas aplicações e impactos no mundo pós-moderno. “A informação comporta um elemento de sentido. É um significado transmitido a um ser consciente por meio de uma mensagem inscrita em um suporte espaço-temporal: impresso, sinal elétrico, onda sonora etc.” (LE COADIC, 2004, p.8).

Ramalho; Nidotti; Fujita, (2007, p. 8) destacam que.

> Utilizando como exemplo um processo de recuperação de informação no âmbito da área da ciência da informação é necessário levar em consideração os componentes semânticos inerentes a tal processo, no entanto, de acordo com o enfoque da área da ciência da computação, observa-se que os tradicionais 'motores de busca', baseiam-se exclusivamente na recuperação de dados, não levando em consideração as semânticas contidas nas páginas *Web*, recuperando apenas sequências de caracteres que satisfaçam determinadas condições de busca.

A larga expansão da *internet* gera muitas informações em forma de opiniões em seus diversos canais: fóruns, comunidades, redes sociais etc. Essa popularização da *internet*, por sua vez, gera um grande volume de informação a cada instante, e as organizações, em geral, não conseguem acompanhar no mesmo ritmo o que os usuários estão comentando sobre as mesmas. No entanto, percebeu-se que ao analisar essas informações, as organizações poderiam ter a vantagem de conhecer as opiniões dos usuários sobre seus serviços ou produtos fornecidos a partir de dados das redes sociais (GOMES, 2013).

Conhecer a opinião dos outros é parte importante no processo decisório de muitos indivíduos (PANG; LEE, 2008). Indurkhya e Damerau (2010) citam que as opiniões são tão importantes que, onde quer que se queira tomar decisões, as pessoas querem ouvir a opinião de

outros. Isso não é uma verdade apenas para as pessoas, como também para as organizações, afinal, conhecer a opinião dos clientes acerca dos seus produtos e serviços é de grande valia para as organizações.

Entretanto a necessidade de descobrir a opinião das pessoas em um processo de tomada de decisão não é recente. Osiek (2014) cita um artigo do crítico E. T. A. Hoffmann publicado no jornal *"Allgemeine Musikalische Zeitung"* em 1810, no artigo, utilizando figuras de linguagem, o crítico dizia:

> "Raios radiantes cortam a noite profunda dessa região, e nos fazem conscientes das sombras gigantes que, em um movimento de ida e volta, fecham-se sobre nós, destruindo tudo dentro da gente exceto a dor da expectativa interminável – essa expectativa que afunda e sucumbe com o prazer que emerge no meio de sons tão jubilantes. Somente através dessa dor, que enquanto consome, mas não destrói o amor, a esperança e a alegria, tenta explodir em nosso peito através de um estrondoso grito proveniente de todas as nossas paixões, podemos viver como espectadores cativados desses espíritos".

Esse texto repercutiu positivamente para o sucesso da, outrora[1] mal recebida pela crítica, "Sinfonia no. 5 em C menor, Op. 67" de Ludwig van Beethoven. A crítica foi importante para Beethoven no século 19 e permanece importante hoje em qualquer atividade (OSIEK, 2014).

Como podemos observar, opiniões são importantes para todas as atividades humanas porque influenciam fortemente as escolhas (LIU, 2012). Empresas e organizações necessitam saber da opinião pública de seus produtos e serviços, bem como possíveis consumidores querem saber a opinião de outras pessoas antes de adquirir tal produto (ibid., 2012).

O campo de estudo que analisa as opiniões e sentimentos das pessoas em textos chama-se análise de sentimento (MEDHAT et al., 2014). O interesse da indústria e da academia neste campo de estudo é, em parte, devido ao seu potencial de aplicações, tais como: *marketing*, relações públicas e campanhas políticas (FREITAS; VIEIRA, 2015). Empresas e organizações podem estar interessadas nas seguintes perguntas:

- O que as pessoas pensam sobre seus produtos, serviços etc.?
- O quão positivo (ou negativo) as pessoas pensam sobre seus produtos?

[1] Um ano e meio antes

- Como as pessoas preferem seus produtos?

Entretanto, analisar os valores positivos e negativos das palavras de um texto e encontrar a sua polaridade não é uma tarefa trivial (PANG; LEE; VAITHYANATHAN, 2002), pois, a classificação da intensidade de sentimento de uma palavra deve ser feita por mais de uma pessoa para não obter resultados errôneos.

Osiek (2014) destaca que não é fácil para um indivíduo consumir um texto como o de Hoffmann e concluir se sua avaliação foi positiva, negativa ou neutra. Tanto a necessidade no processo decisório quanto a dificuldade mencionada motivam o conjunto grande de pesquisas, mais de 7.000, segundo Feldman (2013 apud ibid., 2014), na área de análise de sentimento ou mineração de opinião.

Em adição às dificuldades inerentes ao próprio ser humano existem algumas adicionais difíceis de serem interpretadas pelo computador, sumarizadas por (GRIMES, 2008) da seguinte forma:

> "O desafio tem sua raiz na enorme variabilidade e sutilezas da linguagem escrita e falada: o significado que os seres humanos pegam imediatamente do contexto é muito difícil para o computador detectar. Como pode o *software* discernir de forma confiável fatos e sentimentos a luz não somente de abreviações, má ortografia, regras gramaticais feridas, mas também na existência do sarcasmo, ironia, expressões idiomáticas, gírias e, bem, a personalidade?"

1.1. MOTIVAÇÃO E JUSTIFICATIVA

Em nível mundial há um grande aumento no número de usuários da *internet* e, em consequência, também há uma disponibilidade de grandes quantidades de informações novas na *web*. Grande parte da informação contida na *web* hoje é composta por opiniões e isso se deve principalmente ao advento de ferramentas tais como *blogs*, redes sociais (*Facebook* [2], *Instagram* [3], *Twitter* [4] etc.) e fóruns que permitem aos usuários leigos em informática expressarem suas opiniões na *internet* de forma "anônima".

[2] Disponível em: http://www.facebook.com/
[3] Disponível em: http://www.instagram.com/
[4] Disponível em: http://www.twitter.com/

Olson (2014), no texto a seguir, ajuda a ilustrar a ordem de grandeza do número de pessoas interagindo na *World Wide Web* (WWW).

> "Os 470 milhões de usuários já retiraram US$ 33 bilhões em receita das operadoras com SMS que ficaram ricas cobrando taxas gordas por texto. O *WhatsApp*[5] é gratuito no primeiro ano e depois passa a cobrar U$ 1,00 por ano. Sem propaganda, sem cartazes, sem "*upgrade*" para versões "*premium*". Nas últimas discussões *Zuckerberg*[6] prometeu aos fundadores do *WhatsApp* "pressão zero" para fazer dinheiro, dizendo, - Eu adoraria que vocês conectassem de 4 a 5 bilhões de pessoas nos próximos 5 anos." (OLSON, 2014)

Alguns dados do *Facebook* ampliam a ilustração da magnitude dos números na WWW. Se o *Facebook* fosse um país, seus mais de dois bilhões[7] de usuários formariam a maior população do planeta[8]. Nesse país seriam falados aproximadamente 70 idiomas diferentes.

Todo esse grande contêiner de informações se mostra expressivamente instigante para pesquisadores e empresas. Diversos estudos, pesquisas e desenvolvimento surgem com o intuito de explorar esse potencial que representa, atualmente, o maior bem e poder da humanidade: a informação.

Seu grande potencial se deve, não somente ao continuo crescimento da *internet*, mas também dos mecanismos de buscas.

Estimativas apontam que 85% das informações comerciais estão sob a forma textual (TEXTMININGNEWS, 2011). Fica evidente que tanto a iniciativa pública, quanto à privada, tem grande interesse em explorar tais informações, principalmente, no que tange às opiniões dos consumidores.

Segundo Pang e Lee (2008), cerca de 80% dos usuários já realizaram, ao menos uma vez, pesquisa na Internet sobre algum produto e 20% faz isso em um dia típico, com o intuito

---

[5] Disponível em: http://www.whatsapp.com/
[6] *Mark Zuckerberg* - Fundador do *Facebook*
[7] Dados extraídos de https://g1.globo.com/tecnologia/noticia/facebook-atinge-os-2-bilhoes-de-usuarios.ghtml em novembro de 2017
[8] Ficando à frente da China que possui 1.372.470.000 habitantes. Dados obtidos em https://pt.wikipedia.org/wiki/Lista_de_países_por_população em novembro de 2017

de obter auxílio na tomada de decisão para a compra ou não de um determinado produto ou serviço.

Para Liu (2012) o crescimento do interesse da indústria em relação às opiniões das pessoas sobre os seus produtos e serviços, juntamente com o grande aumento de informações a cada momento disponível na *internet*, culminou no rápido crescimento de técnicas de Análise de Sentimentos. Isto por sua vez impacta diretamente em outras áreas como ciências políticas, econômicas, sociais e como estas são afetadas pelas opiniões de pessoas.

Contando com essas informações disponíveis e com técnicas de análise de sentimentos, este trabalho trata da criação de um protótipo de uma ferramenta *web*, onde o usuário entra com o nome de uma empresa, produto ou serviço, podendo ser na forma de uma palavra ou expressão, e o sistema é capaz de vasculhar na imensidão de dados disponíveis, coletando textos não estruturados e realizando um cálculo probabilístico para identificar qual o sentimento dominante em cada texto.

O *software* será capaz de pesquisar em *blogs*, sites de notícias, e redes sociais, como o *Twitter*, através de robôs de busca (*crawlers*), para trazer os dados necessários para a aplicação.

A motivação deste trabalho está na crescente demanda por ferramentas que processem opiniões, tanto no meio acadêmico quanto no meio corporativo. As empresas precisam avaliar as opiniões sobre seus produtos e serviços, bem como os consumidores precisam de uma ferramenta automatizada para lhes auxiliar na tomada de decisão, haja vista que há um número praticamente ilimitado de maneiras que as pessoas podem expressar suas opiniões (LIU, 2012).

1.2. TEMA E DELIMITAÇÃO

O advento das mídias sociais por meio da *web* tornou fácil a interação entre as pessoas, a publicação de conteúdos e a exposição de opiniões, contribuindo para a *web* se tornar um grande repositório de dados, principalmente em estruturas no formato de texto (PANG; LEE, 2008). A *internet* vem sendo frequentemente utilizada como um meio de troca de informações e opiniões que apresentam um impacto relevante na nossa vida cotidiana (MIAO et al., 2009).

Milhões de mensagens escritas em vários idiomas são enviadas todos os dias, mensagens essas que contêm informações úteis e que poderiam ser usadas para muito mais do que apenas

comunicação. Extrair informações dessas mensagens através de um processo automático pode ser muito útil (DUARTE, 2013).

Estes comentários podem vir a influenciar ou colaborar em tomadas de decisões. Nesse contexto, algumas organizações também estão interessadas neste conteúdo, pois estão percebendo que estas revisões *online* podem trazer algum benefício, como exemplo, inferir em estratégias empresariais (ESULI, 2008).

A definição do problema da análise de sentimentos dada por Liu (2012) nos permite determinar estruturas em um texto complexo de forma didática, o que viabiliza uma compreensão mais clara da área de estudos. Em sua obra, Liu (2012) explora o seguinte exemplo (os números entre parênteses identificam as frases contidas no texto):

> "(1) Eu comprei um *iPhone* alguns dias atrás. (2) Ele parecia ser um ótimo celular. (3) O *touchscreen* era realmente bom. (4) A qualidade de voz era clara também. (5) Porém, minha mãe ficou furiosa comigo por não ter avisado a ela antes de ter feito a compra dele. (6) Ela também achava que o celular era muito caro e queria que eu o devolvesse para a loja."

A questão-chave neste ponto é: o que nós queremos minerar ou extrair dessa passagem? Liu (2012) mostra que na passagem existem diversas opiniões, sejam positivas como as expressas nas sentenças (2), (3) e (4), ou negativas, como nas sentenças (5) e (6). Ele percebe que todas as opiniões possuem alvos. O alvo da opinião na sentença (4), por exemplo, é a qualidade da voz no *iPhone*. E por último, vale ressaltar que todas as opiniões possuem um dono. O dono das opiniões nas sentenças (2), (3) e (4) são o próprio autor da passagem (o "eu").

Nesse contexto, esse trabalho se delimita a analisar a capacidade de predição de um modelo não supervisionado para a análise de sentimentos em dados não estruturados (textos) para que posteriormente um sistema para apoio a tomada de decisões seja desenvolvido com base nesse modelo estudado (*SentiWordNet*).

Juntamente com este software e a instantaneidade e facilidade de acesso da internet, será possível apoiar um indivíduo que, por exemplo, necessita saber qual o impacto de um determinado produto no mercado, onde se pode então, identificar se está na hora de mudar a estratégia, quando se tem um índice desagradável ou acrescentar mais recursos, pelo fato de ter um retorno positivo e alavancar ainda mais suas vendas e a popularidade de sua empresa.

Nesse contexto Stritensky, Stranska e Drabik (2015) defendem que a aplicação de uma análise aprofundada dos dados das mídias sociais e o fornecimento de relatórios prontos para uso em decisões gerenciais são especialmente úteis para definir a estratégia de comunicação e as ações para momentos de crise.

1.3. OBJETIVOS

1.3.1. Objetivo Geral

Esta monografia tem como objetivo principal analisar a capacidade de predição de um modelo não supervisionado para análise de sentimentos em textos (*SentiWordNet*), modelo este que servirá como base para o desenvolvimento de um sistema de apoio a tomada de decisão.

1.3.2. Objetivos Específicos

Para alcançar o objetivo geral do trabalho, percorreremos os seguintes passos:

a) Levantamento bibliográfico para verificar qual é o estado da arte do tema abordado;

b) Analisar a eficiência do modelo escolhido testando-o em alguns contextos distintos através de métricas consolidadas e aceitas na literatura como, por exemplo, acurácia, precisão, revocação, medida F1 e macro F1; e

c) Por fim, desenvolver um protótipo para um sistema *web* de apoio a tomada de decisão baseado no modelo estudado e nos testes realizados.

1.4. RELEVÂNCIA E APLICABILIDADE

Ao longo do tempo os computadores e dispositivos eletrônicos tinham de receber ordens do que fazer. Geralmente, eram programados por desenvolvedores de softwares, antes que eles pudessem completar uma tarefa. Mas nos dias atuais o cenário tem se mostrado um pouco diferente: os computadores estão começando a aprender como executar várias tarefas baseados em experiências, o que se parece muito com o desenvolvimento das habilidades cognitivas dos seres humanos.

Entretanto, diferentemente dos humanos, esses sistemas inteligentes de computação vão reter e lembrar tudo o que eles aprendem.

Partindo dessa premissa, podemos imaginar as possibilidades que passarão a ter cada indivíduo com o acesso a um assistente computadorizado, que por sua vez, vai sempre estar preparado para oferecer informações e orientações úteis com o mais alto nível de relevância nos mais variados segmentos e áreas do conhecimento. Essa é a essência da era da inteligência artificial (IA).

Diante desse panorama, é possível destacar a importância na análise de opiniões disponíveis na rede, pois compreender o que o consumidor, cliente, usuário está dizendo é fundamental para conquistá-lo. Na busca por novas estratégias, para o auxílio na tomada de decisões e para se alcançar vantagem competitiva surge uma nova possibilidade: saber realmente, de maneira probabilística e comprobatória o que as pessoas estão dizendo, pedindo ou querendo.

Portanto a IA, mais especificamente a análise de sentimentos, está preocupada em analisar se um determinado texto está falando algo positivo ou negativo de uma determinada entidade. Neste contexto surgiram estudos com o objetivo de procurar soluções e construir sistemas para facilitar esse processo, o qual é responsável por varrer inúmeros *websites*, buscar opiniões e apresentar os resultados de forma clara para facilitar a decisão por parte do usuário.

1.5. METODOLOGIA

Este trabalho contempla pesquisa de natureza qualitativa-quantitativa, bibliográfica, experimental e interdisciplinar (BOGDAN et al., 1994) em torno do tema Análise de sentimentos auxiliando o processo de tomada de decisão: uma abordagem não supervisionada com o recurso léxico *SentiWordNet*.

Partindo desta premissa, o trabalho está divido em três etapas.

A primeira parte do trabalho está fundamentada na perspectiva cienciométrica de análise das publicações científicas produzidas pelos autores brasileiros e estrangeiros e no conceito de que as publicações científicas são as expressões de pessoas ou grupo trabalhando em uma frente de pesquisa, portanto é possível dizer alguma coisa sobre as relações entre os pesquisadores a partir dessas publicações. (MACIAS-CHAPULA, 1998)

A cienciometria é um segmento da sociologia da ciência que tem como objetivos de estudo as disciplinas, assuntos, áreas e campos científicos, analisando os fatores que diferenciam as subdisciplinas (revistas, autores, documentos, como os cientistas se comunicam) utilizando métodos de análise de conjuntos e correspondências, objetivando identificar os domínios de interesses, concentração de assuntos e compreender a comunicação entre os cientistas. (MACIAS-CHAPULA, 1998).

Santos (2003, p.134) destaca algumas premissas da cienciometria.

> Uma obra científica é produto objetivo da atividade intelectual criativa. Num contexto cientifico, uma publicação é uma representação da atividade de pesquisa de seu autor. O maior esforço deste autor é de persuadir os pares de que sua descoberta, seus métodos e técnicas são particularmente pertinentes. O modo de comunicação escrita fornecerá, portanto, todos os elementos técnicos, conceituais, sociais e econômicos que o autor busca afirmar ao longo de sua argumentação. A atividade de publicação científica é uma eterna confrontação entre as reflexões intrínsecas do autor e os conhecimentos que ele adquiriu pela leitura do trabalho dos outros.

Queiroz e Noronha (2004) observam que o avanço da ciência da informação e das ciências em geral se dá pela constante elaboração de novas pesquisas e pela concretização e divulgação dos resultados que se processam em diferentes tipos de suportes.

Sendo assim, as técnicas cienciométricas são importantes para, entre outras atividades, identificar tendências e o desenvolvimento do conhecimento (SPINAK, 1998).

Existem diversas formas de estudos cienciométricos, dentre elas, podemos destacar a análise de coautoria e os estudos de citação. Ao longo do tempo os estudos de citação têm-se tornado um valioso instrumento para a avaliação dos mais diversos aspectos do fazer científico como: estabelecer o fator de impacto de publicações periódicas, verificar a produção e visibilidade de autores, identificar o impacto dos autores através da quantificação de citações, determinar e/ou identificar frentes de pesquisa etc. Trata-se de análises quantitativas que permitem identificar a disseminação da produção científica (LÓPEZ YEPEZ, 2003).

Deste modo, para a primeira parte no desenvolvimento do trabalho, percorremos os seguintes passos.

a) Levantamento bibliográfico

Etapa indispensável para o trabalho, afinal, nos possibilitou verificar qual é o estado da arte do tema em que desenvolvemos nossas considerações, e mais, indicar qual a posição dos principais autores relacionados ao tema.

b) Leitura

O ponto alto da pesquisa foi a leitura, processo cognitivo de compreensão do texto que fomenta o surgimento de novos textos a partir de instigamento de ideias e da formulação de outros questionamentos que surgem ao longo do processo.

Coube a leitura fornecer subsídios para conectar essas informações, de modo que a pesquisa conseguisse alcançar os seus objetivos.

c) Anotação e síntese da leitura

O fichamento foi uma etapa essencial, tanto para garantir a segurança no desenvolvimento do texto, quanto para maximizar o aproveitamento da leitura.

d) Desenvolvimento do texto

O texto da pesquisa é a consolidação das etapas anteriores, que se somam ao principal componente do trabalho, a opinião do pesquisado sobre o assunto tratado. É aqui que se revela como refletimos sobre os conceitos e os apropriamos para os objetivos visados pela pesquisa.

e) Digitação, normalização e revisão do texto

A digitação e normalização são etapas operacionais, cuja finalidade é adequar o trabalho às normas e os padrões vigentes para a formatação de uma pesquisa científica.

A segunda etapa do trabalho consistirá de dois experimentos realizados com o intuito de analisar a eficiência de um modelo não supervisionado para análise de sentimentos. Nesses experimentos foram analisados contextos específicos, utilizando como fonte de dados o *Twitter*. Para tal, foi desenvolvido um *crawler* simples para a coleta dos *tweets* e, também, uma aplicação para fazer a classificação da polaridade (i. g. positiva, negativa ou neutra) das mensagens. Para o experimento foi utilizado o recurso léxico *SentiWordNet* e os contextos analisados foram: primeiro o das denúncias apresentadas pelo então procurador-geral da república, Rodrigo Janot, contra o atual presidente da república, Michel Temer. E o segundo, o julgamento em segunda instância do ex-presidente Luiz Lula em janeiro de 2018.

A terceira etapa deste trabalho foi norteada pelos objetivos propostos, ou seja, a proposta de uma ferramenta capaz de detectar e classificar automaticamente a polaridade de um texto, assim auxiliando o usuário na tomada de decisão.

1.6. TRABALHOS RELACIONADOS

Análise de Sentimentos (PANG; LEE, 2008) tem uma longa história em Processamento de Linguagem Natural. Desde então, as potenciais aplicações da Análise de Sentimentos, de fato, são inúmeras e utilizadas de forma interdisciplinar tais como previsões do mercado de ações, política, análise nas redes sociais, interação homem máquina etc.

Por Exemplo, Li et al. (LI et al., 2014) implementaram um *framework* genérico para predição dos preços das ações e testaram em seis diferentes abordagens de análise. Eles usaram o dicionário psicológico de Harvard e o dicionário de sentimento financeiro Loughran-McDonald para construir um espaço de sentimento. Artigos de notícias em formato textual foram então quantitativamente mensurados e projetados para tais espaços de sentimentos. A performance foi comparada empiricamente a diferentes níveis de classificação de mercado.

Rill et al. (RILL et al., 2014) propuseram um sistema designado para detectar emergentes tópicos políticos no *Twitter* antes de outros canais de informação. Para as análises, os autores coletaram aproximadamente 4 milhões de *tweets* antes e durante as eleições de 2013 na Alemanha. De abril a setembro de 2013. Os autores compararam seus resultados com o *Google Trends* e perceberam que os tópicos apareciam antes no *Twitter* que no *Google Trends.*

Ohana e Tierney (2009) apresentaram, em seu estudo, os resultados que eles obtiveram ao aplicar o recurso léxico *SentiWordNet* para o problema de classificação automática de sentimentos em *reviews* de filmes. Sua abordagem era composta pela contagem da pontuação positiva e negativa de cada termo para determinar a orientação semântica, e uma melhoria foi apresentada a partir da construção de um conjunto de dados com características relevantes usando o *SentiWordNet* como recurso e em seguida aplicando-o a um classificador de aprendizagem de máquina.

Tem surgido uma grande quantidade de pesquisas em Análise de Sentimentos em linguagens diferentes do inglês através da aplicação de recursos bilíngues e técnicas de traduções automáticas para aplicar técnicas de Análise de Sentimentos em inglês. A abordagem adotada em (HIROSHI; TETSUYA; HIDEO, 2004) usa tradução automática para desenvolver um sistema de Análise de Sentimentos de alta precisão para o Japonês com um baixo custo.

Segundo os autores, o sistema obteve uma precisão de 89%. (MIHALCEA; BANEA; WIEBE, 2007) discutiram métodos para gerar automaticamente um léxico de subjetividade para uma nova língua (o foco deles era o Romeno) a partir de um recurso similar disponível para o idioma Inglês. Eles alcançaram 67.85% de *F-measure* para a classificação da orientação do sentimento de sentenças usando os recursos de subjetividade desenvolvidos para o Romeno. (YAO et al., 2006) propuseram um método para determinar a orientação do sentimento de palavras Chinesas utilizando um léxico bilíngue e eles alcançaram uma precisão e revocação de 92%. (BENAMARA et al., 2007) argumentaram que combinar advérbios com adjetivos são mais úteis para atribuir intensidade de sentimento que adjetivos sozinhos. O melhor algoritmo deles alcançou 47% de acurácia correlacionado com valores atribuídos por humanos comparado à 34% sem usar advérbios.

Hu e Liu (2004), cujo estudo é um dos mais próximos ao nosso, usaram *WordNet* para computar a orientação semântica de avaliação de produtos e tentaram sumarizar as revisões dos usuários extraindo avaliações positivas e negativas de diferentes características de produtos.

Uma grande quantidade de trabalhos tem se direcionado a analisar sentimentos no *Twitter* seguindo diferentes abordagens. (BARBOSA; FENG, 2010) argumentaram que usar *n-grams* em *tweets* pode impedir uma boa performance do classificador por causa da grande quantidade de palavras pouco comuns no *Twitter*. Em vez disso, eles propuseram utilizar recursos do *microblogging* tais como *re-tweets, hashtags e emoticons.* Eles descobriram que usar esses recursos para treinar o SVMs melhoraram a acurácia da classificação do sentimento em 2.2% comparado com o SVMs treinado apenas com *unigrams*. Uma descoberta similar foi reportada em (KOULOUMPIS; WILSON; MOORE, 2011). Eles exploraram os recursos do *microblogging* incluindo *emoticons,* abreviações e a presença de intensificadores tais como letras todas em maiúsculas e a repetição de caracteres para a classificação de sentimentos no *Twitter*.

Outros trabalhos também utilizaram o *Twitter* como fonte de dados, com em (JANSEN et al., 2009), (GO; BHAYANI; HUANG, 2009), (PAK; PAROUBEK, 2010), (O'CONNOR et al., 2010), (TUMASJAN e.t al., 2010), (BIFET; FRANK, 2010), (DAVIDOV; TSUR; RAPPOPORT, 2010), (BOLLEN; MAO, 2011) e (KHAN; BASHIR; QAMAR, 2013). Um dos grandes desafios da análise de sentimentos em redes sociais, muito provavelmente, é a dificuldade de se usar Processamento de Linguagem Natural em textos de baixa qualidade, com muitas gírias, abreviações e erros ortográficos.

Vários outros trabalhos tiveram como objetivo desenvolver um sistema para resolver o problema da classificação de sentimento. Os autores utilizam, em seus sistemas, várias técnicas diferentes para alcançar o resultado mais satisfatório possível. Dentre esses sistemas podemos destacar o *Sentiment Analyzer* que foi desenvolvido em (SALMI, 2015).

O *Sentiment Analyzer* utiliza o *SentiWordNet* para fazer a classificação de sentimentos contidos em textos. A proposta principal desse projeto é comparar a versão em português com a versão em inglês do mesmo *tweet.* Com isso, o autor tinha o intuito de verificar como o sistema se comportava com frases de mesmo contexto em idiomas diferentes.

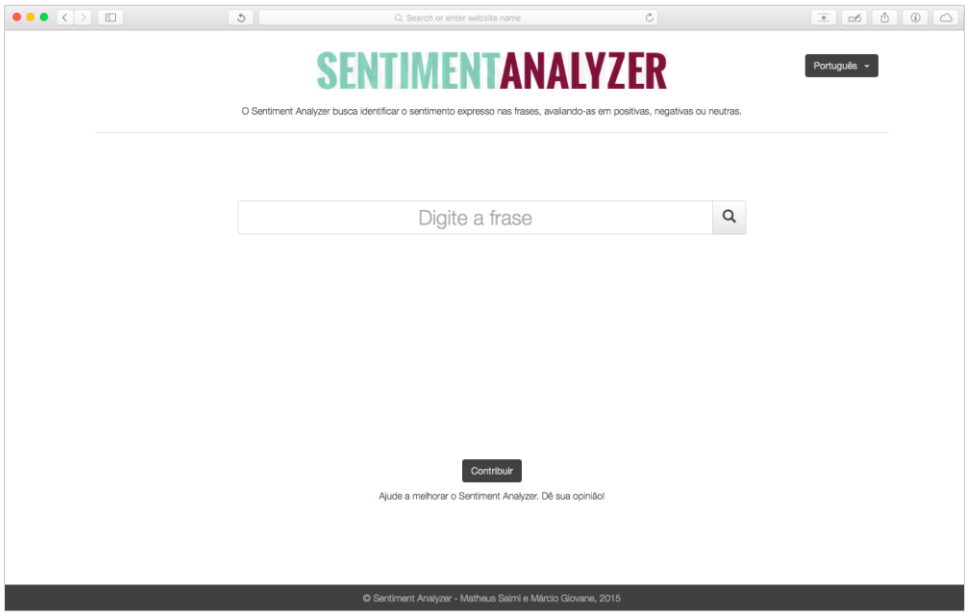
Para os testes, foram selecionados vários *tweets* durante as eleições presidenciais brasileira do ano de 2014. As coletas foram realizadas no período de julho a novembro do mesmo ano, os *tweets* foram selecionados de forma aleatória.

Objetivando realizar o teste com a base em inglês, segundo o autor, os *tweets* foram manualmente traduzidos para a língua, visando transparecer o mesmo conjunto de ideias, mesmo que por palavras de diferente intensidade semântica do idioma de origem, nesse caso, o português.

O autor ainda ressalta que a criação da base de *synsets* e treinamento do sistema em português foi conduzida visando os *tweets* selecionados. Uma vez que treinar uma massa de dados maior se tornaria inviável devido ao tempo e porque, ao final, o sistema seria capaz de classificar poucas ou quase nenhuma frase (SALMI, 2015).

Abaixo temos algumas telas do sistema *Sentiment Analyzer*. A Figura 1 mostra a tela inicial do sistema do sistema, onde é disponibilizado um campo para a pesquisa do termo que se deseja inferir o sentimento. Já na Figura 2 mostra a tela com o resultado da pesquisa. Nela é possível observar a intensidade do sentimento que pode ser: Muito Positivo, Pouco Positivo, Positivo, Muito Negativo, Pouco Negativo, Negativo e Neutro.

Figura 1- Tela Inicial do *Sentiment Analyzer*



Fonte: Salmi (2015)

Figura 2 - Tela de análise de sentimento de frases

Fonte: Salmi (2015)

O autor destaca que no *Sentiment Analyzer*, foi explorado uma dinâmica que envolve feedback de usuários sobre os synsets previamente cadastrados. Isso garante uma flexibilidade do sistema em questão de adaptação, já que os conceitos são refinados com o tempo, mas, paralelamente, proporciona uma barreira para elaborar a base de conhecimento, uma vez que depende de um alto número de votações, ou seja, origina-se de demasiado trabalho humano (SALMI, 2015).

1.7. ESTRUTURA DO TRABALHO

Esta monografia encontra-se dividida em sete capítulos, estruturado da seguinte forma: no primeiro capítulo fazemos uma breve introdução do tema, abordando de uma maneira geral o trabalho apontando o contexto no qual está inserido, além de mostrar os objetivos gerais e específicos aos quais este trabalho é direcionado.

O segundo capítulo apresenta os principais conceitos e terminologias aplicados a análise de sentimentos, descreveremos sobre suas origens e a aplicabilidade do tema da pesquisa, apresentaremos alguns métodos utilizados para se inferir sentimentos em textos, suas aplicações, bem como os desafios que dificultam o processo de mineração de opinião.

O terceiro capítulo tem por objetivo apresentar o estado da arte em Mineração de Texto. Apresentando alguns conceitos e técnicas dessa área e mostrando de uma maneira detalhada os passos necessários para se realizar a extração de informação em documento textual.

O quarto capítulo apresenta dois experimentos realizados no contexto da política para validar o modelo escolhido para se realizar a classificação de sentimentos (*SentiWordNet*) bem como percorrer o processo de coleta dos dados e as métricas necessárias para a análise.

Finalmente, no quinto capítulo são apresentados e discutidos os resultados, obtidos com a execução do classificador, para ambos os experimentos através de gráficos e análise.

No sexto capítulo é feita uma breve apresentação do sistema *web* desenvolvido a partir do modelo estudado e dos testes realizados.

Por fim, no sétimo capítulo é sumarizada as principais conclusões originadas deste trabalho, além de apresentar possíveis direções de pesquisa para trabalhos futuros.

2. UMA PANORÂMICA NA ANÁLISE DE SENTIMENTOS

Na literatura é possível encontrar diversos trabalhos relacionados à análise de sentimentos ou *opinion mining* (mineração de opinião) de frases extraídas das redes sociais. O objetivo desta seção é apresentar os principais aspectos conceituais, discutindo os estudos acadêmicos encontrados nos últimos anos.

2.1. ORIGEM E CONCEITOS DESSA ÁREA

Embora seja difícil precisar o início das pesquisas nessa área, credita-se sua origem à pesquisa em extração de crenças (o que A conhece a respeito de B) e "ponto de vista" de texto, utilizando técnicas de IA simbolista no processamento de linguagem natural (WILKS; BIEN, 1984).

A análise de subjetividade é definida por Wiebe (1994) como a expressão linguística de opinião, sentimento, emoção, avaliação, crenças e especulações de alguém. Em sua definição, o autor estava inspirado pelo trabalho do linguista Ann Banfield (BANFIELD, 1982), que define como subjetividade as frases que se caracterizam do ponto de vista (USPENSKY, 1973) e que apresentam um estado privado (QUIRK, 1985) (i. e. estados que não estão abertos à observação ou verificação de objetividade) de um experimentador. Subjetividade é o oposto de objetividade, que é a expressão de fatos.

O ano de 2001 marcou a conscientização generalizada do problema. Três fatores estão por trás desse movimento (PANG; LEE, 2008).

1. A disponibilidade de banco de dados e o desenvolvimento de sites agregadores de opinião; e
2. O surgimento de métodos de aprendizado de máquina no processamento de linguagem natural e recuperação da informação;
3. A realização de que se trata de um problema que impõe desafios para a pesquisa com aplicações na área comercial e de inteligência.

No estudo realizado por Pang e Lee (2008), o termo "análise de sentimentos" é descrito com o mesmo significado de mineração de opiniões, que é um ramo da mineração de textos com foco, não na classificação de tópicos, mas na classificação de acordo com os sentimentos, ideias e opiniões das pessoas a respeito de um assunto.

A análise de sentimentos tem como objetivo determinar a intensidade de sentimentos e a polaridade das frases capturadas da *web* (PANG; LEE; VAITHYA-NATHAN, 2002). A polaridade de uma frase representa as características positivas, negativas ou neutras da frase. Os sentimentos expressam o grau de intensidade positiva ou negativa de uma frase, possuindo uma escala que pode variar, por exemplo, de -5 a +5.

Segundo Xu (2010), Mineração de Opinião ou Análise de Sentimentos são áreas que abrangem técnicas computacionais que procuram entender opiniões e sentimentos em textos, analisando uma grande quantidade de dados com intuito de auxiliar as pessoas em tomada de decisões. Para Pang e Lee (2008), é a área que estuda o tratamento computacional de opinião, sentimento e subjetividade em texto. Jun-Ping (2009) cita como uma recente disciplina que envolve recuperação da informação, processamento de linguagem natural e linguística computacional. Enquanto Mukherjee (2012) diz que análise de sentimentos consiste em uma atividade que envolve Processamento de Linguagem Natural (PLN) Extração de Informações e que visa obter o sentimento das pessoas que são expressos em comentários positivos ou negativos, perguntas e pedidos em documentos escritos através da análise de uma grande quantidade desses elementos.

Para Ribeiro at. al. (2015) o principal objetivo da análise de sentimentos é definir técnicas automáticas capaz de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão.

Embora a opinião seja um conceito muito amplo, a análise de sentimento foca, principalmente, sentimentos positivos e negativos. (MOREO, et al., 2012). Ela é usada para extrair opiniões, sentimentos e subjetividade em textos não estruturados, ou seja, para identificar se as expressões indicam um parecer positivo (favorável) ou negativo (desfavorável) para o assunto (PANG; LEE, 2008).

2.2. ALGUNS MÉTODOS UTILIZADOS NA MINERAÇÃO DE OPINIÃO

Nessa seção mostraremos alguns estudos e métodos criados para se realizar a análise de sentimentos em textos. Aqui serão mostradas técnicas que usam basicamente dicionários léxicos (a definição formal de dicionário léxico será mostrada na seção 4.1).

Léxicos são utilizados para o problema da análise de sentimentos para tentar sanar o problema dos classificadores que, em geral, reside no alto custo de treinamento. Trabalhos como os realizados em (DENECKE, 2008), (OHANA; TIERNEY, 2009), (TABOADA et al. 2011), (NIELSEN, 2011) e (CAMBRIA et al., 2016), fazem uso de léxicos como base para a classificação.

Ribeiro et al. (2015) ressaltam que essas abordagens têm sido utilizadas como métodos de prateleira (*off-the shelf*), isto é, pesquisadores e demais usuários interessados em aplicar a análise de sentimento para algum propósito específico escolhem alguma das soluções disponíveis e aceitas na literatura e aplicam para o fim desejado.

2.2.1. *SentiWordNet*

SentiWordNet (SWN) (ESULI; SEBASTIANI, 2006) é uma ferramenta muito utilizada em mineração de opinião, e é baseado no dicionário léxico *WordNet*[9] (MILLER, 1995). Esse dicionário agrupa adjetivos, verbos e outras classes gramaticais em conjuntos chamados *synset*. O SWN associa a cada *synset* do *WordNet* três valores de pontuação que indicam o sentimento de um texto: positivo, negativo e objetivo (neutralidade). Cada pontuação é obtida utilizando um método de aprendizagem de máquina semi-supervisionada, e variam de 0 a 1, com soma igual a 1. Nessa monografia utilizamos a versão 3.0 do SWN, disponível em (SENTIWORDNET, s. d.).

"A SWN é distribuída sob a licença *Attribution-Share Alike 3.0 Unported* (CC BY-AS 3.0). Esta licença permite o uso do SWN em aplicações comerciais, desde que o aplicativo mencione o uso do SWN e SWN é atribuído a seus autores" (SENTIWORDNET, s. d.).

A SWN é uma ferramenta disponível apenas para a língua inglesa e que utiliza a *WordNet* 3.0, uma base de dados léxica dessa língua. Nela, nomes, verbos, adjetivos e advérbios são agrupados em um conjunto de sinônimos cognitivos (*synsets*), cada um expressando um conceito distinto. Essa ferramenta é amplamente utilizada para a língua inglesa e tem apresentado bons resultados.

Cada *synset* é associado a três valores numéricos, Pos(s), Neg(s) e Obj(s) que indicam o quanto é positivo, negativo ou objetivo (neutro) os termos contidos no *synset*. Cada um dos

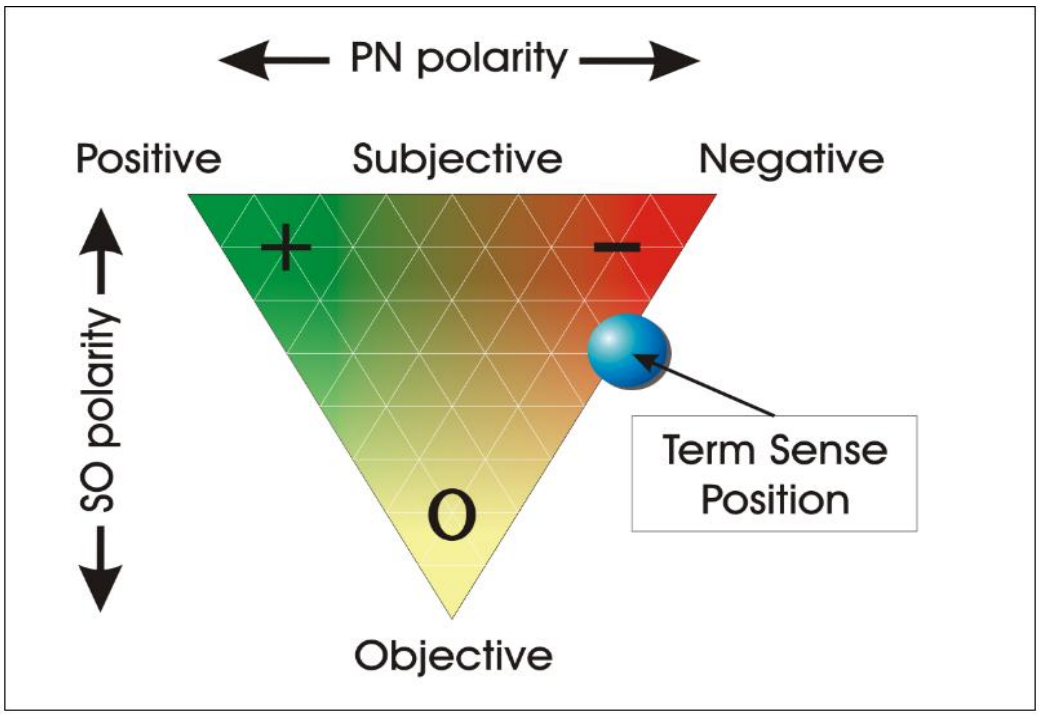
[9]Disponível em: http://wordnet.princeton.edu/

três valores varia no intervalo [0.0, 1.0] e a soma deles é 1.0 para cada *synset*, logo Obj(s) + Pos(s) + Neg(s) = 1.0.

O que motivou a atribuição de valores numéricos para um synset, ao invés de valores diretos a um termo, foi a possibilidade de um mesmo termo ter diferentes sentidos e cada um desses sentidos merece uma pontuação diferente (ESULI; SEBASTIANI, 2006). Na SWN, por exemplo, a palavra "*broken*" está relacionada a dois *synset*, S1 = {*wiped out, impoverished, broken}* que tem como descrição "*destroyed financially or the broken fortunes of the family*" e pontuações Obj(s) = 0.5, Pos(s) = 0 e Neg(s) = 0.5, e S2 = {*broken*} que tem a descrição "*physically and forcibly separated into pieces or cracked or split*" e pontuações Obj(s) = 0.875, Pos(s) = 0 e Neg(s) = 0.125. É possível notar que a SWN tem uma pontuação diferente para diferentes significados de um mesmo termo.

A Figura 1 ilustra a representação gráfica adotada pelo *SentiWordNet* representando as propriedades relacionadas ao *synset*. As bordas do triangulo representam uma das três classificações (positivo, negativo e objetivo) e um ponteiro (*synset position*) aponta para a classificação de maior valor.

Figura 3 - Representação gráfica adotada pelo *SentiWordNet*



Fonte: Esuli e Sebastiani (2006)

Baccianella et al. (2010) citam quatro diferentes versões do SWN que foram discutidas em publicações:

1. SentiWordNet 1.0, apresentado em Esuli e Sebastiani (2006) e publicamente disponíveis para pesquisa;
2. SentiWordNet 1.1, apenas discutido em relatório (ESULI; SEBASTIANI, 2007) que nunca chegou ao estágio de publicação;
3. SentiWordNet 2.0, apenas discutido na tese de Doutorado de Esuli (2008);
4. SentiWordNet 3.0, (BACCIANELLA, et al., 2010) versão utilizada nesse trabalho.

Para a SWN 1.0 e 1.1 foram empregados algoritmos de aprendizagem supervisionados e semi-supervisionados. Já para a versão 2.0 e 3.0 os resultados destes algoritmos semi-supervisionados são adotados com uma etapa intermediária no processo de etiquetagem, que são alimentados para um processo interativo denominado *random-ralk*, atingida a convergência deste processo resulta o SWN 3.0.

A base de dados do léxico SWN é disponibilizada em arquivo de dados (*.txt*). Na Tabela 1 exibimos um exemplo de registros no *SentiWordNet* 3.0.

Tabela 1 - Exemplo da base de dados do *SentiWordNet* 3.0

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	02259239	0.375	0.0	sold#1	disposed of to a purchaser; "this merchandise is sold"
n	00005930	0.0	0.125	dwarf#3	a plant or animal that is atypically small
r	00003294	0.375	0.25	anisotropically#1	in an anisotropic manner
v	00605671	0.0	0.375	brainwash#2	submit to brainwashing; indoctrinate forcibly

Fonte: O autor (2017)

O POS (*part of speech*) identifica a classe gramatical a qual aquela palavra pertence, os valores *PosScore* e *NegScore* correspondem a positividade e negatividade assinalada pelo *SentiWordNet* para o determinado *synset*. O valor de objetividade é dado por: Obj(s) = 1 - (Pos(s) + Neg(s)). A coluna *SynsetTerms* corresponde aos termos separados por espaço, pertencente ao *synset*, com a classe gramatical e número correspondente ao sentido. A coluna *Gloss* descreve o sentido do termo.

2.2.2. *SentiStrength*

O *SentiStrength* é um classificador baseado em léxico, construído por (THELWALL et al., 2010) com a finalidade de detectar sentimentos em textos curtos da língua inglesa. *SentiStrength* combina uma abordagem baseada em léxico com regras linguísticas mais sofisticadas, por exemplo, erros ortográficos, pontuação e uso de *emoticons*. O léxico classifica, de forma automática, até 16.000 textos por segundo com precisão como se fosse efetuado por um humano (SENTISTRENGTH, 2017).

O *SentiStrength*, disponível *online*[10], para cada entrada de texto, classifica as palavras em um intervalo de 1 a 5 denotando palavras positivas, e -5 a -1 para palavras negativas (GARCIA; SCHWEITZER, 2011). Por exemplo, um texto com uma pontuação de 3 e -5 contém sentimento positivo moderado e forte sentimento negativo, respectivamente. A frase "*I love you but hate the current political climate*" é classificada da seguinte maneira: "*I love*[3] *you but hate*[-4] *the current political*", para cada palavra avaliada da frase é extraída uma pontuação mínima de -5 e máxima de 5, pertinente ao seu conteúdo emocional inserido no léxico (SENTISTRENGTH, 2017).

As palavras que aparecem no texto, e por ventura não existam no léxico (SENTISTRENGTH, 2017), não são analisadas e consequentemente não interferem no resultado da classificação final. Além desta classificação, o analisador ainda oferece uma classificação única, ou seja, fazendo uma análise no geral (texto inteiro). A análise é feita pegando o maior positivo e o maior negativo e efetua a diferença.

Outro fator importante do classificador é a possibilidade da análise de textos em diversos idiomas, entre eles o português, mas não com a precisão que se tem para o inglês.

2.2.3. LIWC

O LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta proposta em (PENNEBAKER et al. 2001) e tem o objetivo de analisar os componentes emocionais, cognitivos e estruturais de textos. Isso é feito com base em um dicionário de palavras, que pode ser encontrado em diversas línguas.

> O LIWC é capaz de calcular como as pessoas usam diferentes categorias de palavras através de uma ampla variedade de textos. Seja em e-mails, discursos, poemas ou a transcrição de qualquer diálogo

[10] Disponível em: http://sentistrength.wlv.ac.uk/

> cotidiano, o LIWC permite determinar o grau em que os autores/falantes usam palavras que conhecem emoções positivas ou negativas, referências autênticas, palavras estendidas ou palavras que se referem ao sexo, comer ou religião. O programa foi projetado para analisar de forma simples e rápida mais de 70 dimensões de linguagem através de centenas de amostras de texto em segundos. (LIWC, s. d.)

O LIWC é uma ferramenta de análise textual que realiza uma avaliação de cada uma das palavras contidas em um texto tendo como base um dicionário pré-definido pelo usuário. Essa avaliação consiste na tentativa de relacionar palavras a pelo menos uma das categorias definidas, tendo como base teorias da esfera psicológica, onde cada relação entre uma palavra e uma categoria indica uma maior tendência do usuário a uma determinada personalidade ou atitude. Na versão paga da ferramenta são disponibilizadas para avaliação mais de 100 categorias de palavras.

LIWC é uma ferramenta para análise de texto que estima componentes emocionais, cognitivos e estruturais de um dado texto baseada no uso de dicionários contendo palavras e suas respectivas categorias. A título de exemplo, no LIWC a palavra "*agree*" pertence a 5 categorias: *assent*, *affective*, *positive emotion*, *positive feeling*, e *cognitive process*. Portanto, além de detectar *positive* ou *negative feeling* em um texto, o LIWC também fornece outras categorias de palavras. (ARAUJO; GONÇALVES; BENEVENUTO, 2013)

2.3. APLICAÇÕES DA ANÁLISE DE SENTIMENTOS

Podemos utilizar análise de sentimentos nos mais variados setores do mercado consumidor, tais como avaliação de produtos, descoberta de atitudes e suas tendências de consumidores para o fortalecimento de campanhas de *marketing*, encontrar opiniões acerca de tópicos em alta ou também avaliar filmes etc.

Esta seção tem como objetivo elencar alguns dos principais ramos de atuação da análise de sentimentos.

2.3.1. Análise de Sentimento em Sistemas de Recomendação

Smith (2011) mostra que pessoas tendem a coletar opiniões de sites ou pessoas que compartilham seu ponto de vista. Nesse sentido, (FILHO, 2006) desenvolveu o sistema *e-Recommender*. Este sistema utiliza uma abordagem colaborativa para encontrar usuários com

gostos semelhantes. Ele utiliza, também, uma abordagem de conteúdo para comparar características descritivas dos produtos, para encontrar produtos interessantes ao usuário.

O projeto *MovieLens*[11], analisado em (SILVA, 2014), é uma das iniciativas do grupo de pesquisa *GroupLens*[12] para recomendação de filmes. Este projeto é utilizado como ferramenta de pesquisa para avaliação do usuário de questões relacionadas à Sistemas de Recomendação, como: recomendadores sociais, recomendadores baseados em marcadores (*tagging*) e interface com o usuário.

Quando o usuário faz seu registro no sistema ele recebe primeiro uma lista aleatória de filmes para que o mesmo faça uma avaliação para cada um deles. As avaliações são computadas por meio de uma escala numérica entre 0 e 5 representada por estrelas. Neste caso é possível notas com fração de 0.5 (metade de uma estrela) e quanto mais estrelas possuir melhor é a avaliação dos usuários quanto ao filme. A tarefa de recomendação original do sistema é prever a avaliação de um usuário para um filme antes que o mesmo o assista. Para isso, o sistema compara as avaliações de um usuário com a de outros usuários que tenham realizado avaliações semelhantes para os mesmos filmes. Ao final o sistema identifica as notas dadas pelos usuários semelhantes (vizinhos) e infere a nota que o usuário daria ao assistir ao filme.

2.3.2. Detecção de Mensagens de Ódio

Qualquer ato de comunicação escrita ou falada que busque ofender ou intimidar qualquer cidadão ou grupo deles é chamado discurso de ódio (do inglês, *hate speech*). (MACEDO, 2017)

Em seu artigo, Macedo (2017), cita o projeto "Comunica que Muda"[13]. Este projeto executou um algoritmo computacional para vasculhar o *Twitter*, *Facebook* e o *Instagram* aqui no Brasil em busca de textos com esse teor. O algoritmo foi rodado entre abril e junho de 2016 e nesse curto espaço de tempo, foram detectadas quase 400 mil menções.

[11] Disponível em: http://www.movielens.umn.edu/
[12] Disponível em: http://www.grouplens.org/projects
[13] Disponível em: http://www.comunicaquemuda.com.br/dossie/quando-intolerancia-chega-as-redes/

2.3.3. Análise de Sentimento na Política

Muitos trabalhos foram realizados nessa área. Interessante notar, talvez pela quantidade de informação disponibilizada diariamente na web, que serviram de base para elaborar e testar novos algoritmos.

Nesse trabalho, por exemplo, nas seções 4 e 5, são realizados dois experimentos para testar o desempenho de um modelo para análise de sentimento, e para tal, foi feita uma coleta de aproximadamente 2000 *tweets* relacionados às denúncias do então procurador-geral da república, Rodrigo Janot, contra o atual presidente da república, Michel Temer em 2017 e ao julgamento do ex-presidente Lula no TRF-4 em janeiro de 2018.

O trabalho apresentado em (FRANÇA; OLIVEIRA, 2014) teve como objetivo analisar a polaridade expressa nas mensagens publicadas no *Twitter* durante os protestos que ocorreram no Brasil nos meses de junho, julho e agosto de 2013. Para realizar a análise, foi utilizado o modelo estatístico de aprendizagem *Naive Bayes.*

Oliveira e Bermejo (2017) realizaram um estudo para identificar como a análise de sentimento, baseada em textos extraídos de mídias sociais, pode ser um instrumento de mensuração pública sobre a atuação do governo, de forma a contribuir para a avaliação da administração pública. O aplicativo utilizado para realizar a mineração de opinião foi o *DiscoverText*[14], sua escolha, segundo o autor, foi devido a sua disponibilidade gratuita para fins acadêmicos e obter resultados satisfatórios em outros trabalhos.

2.4. DESAFIOS DA ANÁLISE DE SENTIMENTOS

Em seu trabalho, Subhabrata (2012) aborda os seguintes desafios para a análise de sentimentos:

2.4.1. Sentimentos implícitos e sarcasmos

Sarcasmo é uma forma sofisticada de discurso onde qualquer falante ou escritor diz ou escreve o oposto do que realmente quer dizer. Sarcasmo pode ser estudado em linguística, psicologia e ciências cognitivas (GIBBS, COLSTON, 2007; GIBBS, 1886; KREUZ, CAUCCI, 2007; KREUZ, GLUCKSBERG, 1989; UTSUMI, 2000). No contexto de análise de

[14] Disponível em: http://www.discovertext.com

sentimentos, isso quer dizer que quando alguém diz algo positivo ele/ela realmente quer dizer algo negativo, e vice-versa. Sarcasmo é muito difícil de ser detectado, porém, alguns trabalhos para esta finalidade foram feitos em (GONZÁLEZ-IBÁÑEZ, MURESAN, WACHOLDER, 2011; TSUR, DAVIDOV, RAPPOPORT, 2010).

Liu (2012) afirma que sentenças sarcásticas não são muito comuns em *reviews* de produtos e serviços, mas elas são muito frequentes em discussões *online* e comentários sobre políticos.

As sentenças de um texto podem apresentar sentimentos implicitamente mesmo que não haja presença clara destes através de palavras. Considere o exemplo abaixo:

"Como alguém consegue passar uma tarde inteira na fila do SUS!?"

Percebe-se que essa sentença não carrega explicitamente um teor negativo através das palavras, apesar de ela possuir essa polaridade.

2.4.2. Dependência de domínio

Aqui é preciso compreender que algumas palavras têm sua polaridade modificada de acordo com o domínio no qual se encontram. Analisemos a seguinte frase:

"Leia o livro O Caçador de Pipas".

Essa frase possui um sentimento positivo se considerar o domínio "leitura de livros", tendo em vista que é uma indicação do produto, entretanto, se considerarmos o domínio de "filmes que reproduzem livros" pode haver um teor negativo, tendo em vista que o diretor do filme pode receber o *feedback* negativo de que precisa ler o livro para reproduzi-lo melhor no filme.

O que agrada a alguns não necessariamente agradará a outros. O que é considerado positivo por uns pode ser considerado negativo por outros. Uma avaliação positiva do produto de um fornecedor pode ser considerada negativa por seu concorrente. Civilizações em conflito tem afetividades distintas sobre um mesmo fato. A classificação afetiva de uma opinião é dependente de contexto (OSIEK, 2014).

2.4.3. Expectativas frustradas

Este caso não está relacionado a situações em que o autor insere o contexto em sua mensagem apenas para refutá-lo ao final desta. Considere o seguinte exemplo:

"Nossa, essa viagem deveria ser maravilhosa! Encontramos passagens aéreas baratas, conseguimos um excelente hotel e receberíamos um serviço turístico renomado na região. Porém, tudo acabou indo por água abaixo!".

Apesar de termos palavras agradáveis a maior parte do texto e que indicam um sentimento positivo, o texto na realidade expressa um sentimento negativo em relação a situação, posto que a última sentença é crucial na definição da polaridade da mensagem.

2.4.4. Pragmática

A pragmática, ramo da linguística que estuda a linguagem dentro do domínio de sua aplicação – a comunicação, é um elemento importante a ser detectado, posto que ela é capaz de alterar completamente o sentimento do usuário. Veja os seguintes exemplos:

"Caramba, meu time DESTRUIU o seu no jogo de ontem, hein?!"

"Estou completamente destruído após um dia inteiro de trabalho!"

Percebe-se que o uso de caixa alta no verbo destruir na primeira frase denota um sentimento (positivo). Por outro lado, este mesmo verbo denota um sentimento negativo na segunda frase, significando o mesmo que exausto.

2.4.5. Conhecimento de mundo

É importante que algum conhecimento de mundo seja adicionado aos sistemas que analisam sentimentos. Veja o seguinte exemplo:

"Ela é uma verdadeira bruxa!".

Esta frase retrata um sentimento negativo. Porém, para identificá-lo, é preciso ter um conhecimento de mundo sobre o que representa o termo "bruxa" na frase.

2.4.6. Detecção de subjetividade

Consiste na diferenciação de textos sem opinião. É uma técnica que auxilia na detecção de sentimentos na medida em que pode compor filtros que retiram as mensagens objetivas do conjunto de dados a serem analisados. Veja os exemplos a seguir:

"Comprei o novo *smartphone* da Motorola ontem, o *Milestone 3".*

"Gostaria de comprar um celular que fosse leve".

"Odiei o novo celular que ganhei!".

O primeiro exemplo representa uma frase objetiva. O segundo apresenta uma frase subjetiva, porém o autor da mesma não expressa uma opinião positiva ou negativa em relação a algo. Por fim, a terceira frase é subjetiva e expressa uma opinião negativa em relação ao celular.

2.4.7. Identificação de entidade

Por vezes em uma mesma sentença temos a presença de mais de uma entidade. Frases comparativas são exemplos clássicos desse caso. É importante, portanto, identificar para qual das entidades do contexto a opinião dada é direcionada. Veja os exemplos abaixo:

"O Motorola *Razr* é melhor que o *iPhone* 5".

"O Vasco foi muito superior ao Flamengo no jogo do Campeonato Brasileiro de ontem".

Os exemplos acima representam sentimento positivo em relação a Motorola *Razr* e Vasco e sentimento negativo em relação a *iPhone* 5 e Flamengo.

2.4.8. Negação

A manipulação de negações dentro do contexto de uma frase é um dos grandes desafios da análise de sentimentos. Isso se deve ao fato de que, além de uma negação poder ser expressa de várias maneiras (algumas vezes explícita e outras implicitamente), é preciso que se considere o escopo da negação para saber se ela realmente se caracteriza como uma. Vamos considerar os seguintes exemplos:

"Eu não gosto do Fantástico".

“Eu não gostei do elenco do filme, mas adorei a direção do *Steven Spielberg*.”

“Eu não apenas gostei do elenco do filme, bem como adorei a direção do *Steven Spielberg*.”

O primeiro exemplo representa o método mais simples para identificação de negação, que consiste em encontrar o operador de negação (no exemplo a palavra “não”) na frase e reverter a polaridade desta.

Este método, porém, é muito simples para identificar que a frase não está completamente negativada no segundo exemplo, pois ele não é capaz de identificar o escopo da negação. Neste caso é preciso perceber que o operador de negação “mas” está modificando a polaridade da segunda sentença. O método, portanto, consiste em modificar a polaridade de todas as palavras seguintes a negação até se encontrar outra negação.

O terceiro exemplo representa um caso que os métodos anteriores não conseguem solucionar. A polaridade da primeira sentença não é alterada em vista da presença o operador “não” em vista do termo “apenas” que vem em seguida. Portanto, um método mais robusto deve ser capaz de identificar expressões como “não apenas” e afins.

3. DESCOBERTA DE CONHECIMENTO EM TEXTO (KDT)

O objetivo desta seção é descrever o estado da arte em Mineração de Textos (MT), através de seus conceitos e definições, do detalhamento das etapas de seu processo e da identificação de várias técnicas de descoberta de conhecimento em textos.

3.1. MINERAÇÃO DE TEXTOS – CONCEITOS E DEFINIÇÕES

A mineração de textos tem sua origem relacionada à área de descoberta de conhecimento em textos (*Knowledge Discovery from Text* – KDT), tendo seus processos sendo descritos pela primeira vez em (FELDMAN; DAGAN, 1995, p.112-117), descrevendo uma forma de extrair informações a partir de coleções de texto dos mais variados tipos.

Existem diversos conceitos sobre mineração de texto: "A Mineração de Textos, também conhecida como Descoberta de Conhecimento de Texto refere-se ao processo de extrair padrões interessantes e não triviais ou conhecimento a partir de textos desestruturados" (TAN, 1999, p.1, tradução livre). Moura (2004) descreve a mineração de textos, como sendo uma área de pesquisa tecnológica cujo objetivo é a busca por padrões, tendências e regularidades em textos escritos em linguagem natural.

Já Wives (2002), afirma que a mineração de textos pode ser entendida como a aplicação de técnicas de KDD (*Knowledge Discovery in Database*) sobre dados extraídos de textos. Entretanto, KDT não inclui somente a aplicação das técnicas tradicionais de KDD, mas também qualquer técnica nova ou antiga que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto. Com isso, muitos métodos foram adaptados ou criados para suportar esse tipo de informação semiestruturada ou sem estrutura, que é o texto.

A MT aplica as mesmas funções analíticas da Mineração de Dados (MD) (GOMES, 2013), porém para dados textuais. Segundo Hearst (1999), os dados textuais englobam uma vasta e rica fonte de informação, mesmo em um formato que seja difícil de extrair de maneira automatizada.

Assim como a MD, que obtém informações de banco de dados e/ou outras fontes estruturadas, a MT se alimenta, principalmente, de dados não estruturados ou semiestruturados. Essa é a principal diferença entre os dois métodos segundo Gupta e Lehal (2009).

Seu grande potencial se deve ao crescimento contínuo da *internet* e dos mecanismos de busca que tem sido expressivo nos últimos anos.

Apesar da falta de consenso na literatura atual, o KDT, assim como o KDD, é um processo (metódico) que tem como intenção a obtenção de conhecimento e pode ser dividido em algumas fases.

Afim de melhor contextualizar o leitor com os aspectos relativos ao processo de KDT as próximas seções serão para discorrer sobre esses processos e, por fim, na seção 3.5 será apresentado todo o processo de mineração de texto.

3.2. RECUPERAÇÃO DE INFORMAÇÃO

Recuperação de Informação (RI) é uma área da computação que lida com o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos.

De acordo com Calvin Mooers (1951 apud SARACEVIC, 1996, p. 44), o termo recuperar informação "engloba os aspectos intelectuais de descrição de informações e suas especificidades para a busca, além de quais quer sistemas, técnicas ou máquinas empregados para o desempenho da operação".

Belkin e Croft (1987) definem o processo de RI como um processo de localização de itens de informação que tenham sido objetos de armazenamento, com a finalidade de permitir o acesso dos usuários aos itens de informação, objetos de uma solicitação. A recuperação da informação se dá pela comparação do que se solicitou com o que está armazenado e com o conjunto de procedimentos que esse processo envolve.

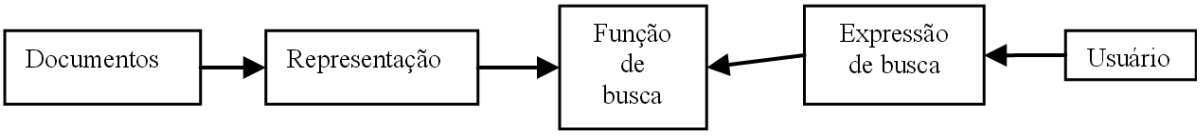
Segundo Salton e McGill (1983), os processos para recuperação de informação necessitam de técnicas que agilizam o armazenamento e acesso aos dados. Estas técnicas envolvem a atribuição de termos a própria dose de identificadores para representar o conteúdo dos documentos na coleção. Esta tarefa, conhecida como indexação (detalhada na seção 3.5.3), pode ser feita automaticamente ou manualmente.

A recuperação da informação é feita através de uma entrada do usuário, ou seja, através de uma consulta para que os documentos relevantes sejam encontrados. Os processos de RI

geralmente se baseiam em buscas por palavra-chave ou busca por similaridade (KAMBER, 2001).

O processo de recuperação da informação está baseado em coleta, representação, armazenamento, organização e acesso por parte dos usuários. De modo geral, o processo de um sistema de informação detém aspectos linguísticos e objetos textuais, portanto necessita de interpretação correta dos elementos envolvidos, o que garante uma recuperação com qualidade. De forma simplificada, a Figura 3, abaixo, representa um modelo básico de sistema de recuperação de informação.

Figura 4 - Modelo simplificado de um sistema de recuperação de informação



Fonte: Baseado em Ferneda (2003)

Portanto, recuperar uma informação consiste em identificar, em um acervo documental, quais os documentos satisfazem total ou parcialmente a uma determinada necessidade de informação do usuário.

3.3. APRENDIZAGEM DE MÁQUINA

Segundo Coppin (2010), aprendizado está diretamente ligado com a inteligência, pois realmente se um sistema é capaz de aprender a exercer determinada tarefa mereça então ser chamado de inteligente.

> Um processo de aprendizagem inclui a aquisição de novas formas de conhecimento: o desenvolvimento motor e a habilidade cognitiva (através de instruções ou prática), a organização do novo conhecimento (representações efetivas) e as descobertas de novos fatos e teorias através da observação e experimentação. Desde o início da era dos computadores, tem sido realizada pesquisas para implantar algumas destas capacidades em computadores. Resolver este problema tem sido o maior desafio para os pesquisadores de inteligência artificial (IA). O estudo e a modelagem de processos de aprendizagem em computadores e suas múltiplas manifestações constituem o objetivo principal do estudo de aprendizado de máquinas. (SANTOS, 2005, p10).

Aprendizado de Máquina (AM) estuda a criação de modelos de algoritmos probabilísticos capazes de “aprender” através da experiência. O aprendizado se dá através de métodos dedutivos para extração de padrões em grandes massas de dados (CHAKRABARTI, 2002). AM tem sido muito utilizado no processo de classificação automática de textos.

No contexto de análise de sentimentos, os algoritmos de Aprendizagem de Máquina são treinados com um conjunto de dados previamente rotulados e, assim, são capazes de classificar uma nova instância de acordo com o conhecimento adquirido (DUARTE, 2013).

Pesquisas de AM estudam o desenvolvimento de métodos capazes de extrair conceitos (conhecimento) a partir de amostras de dados (MITCHELL, 1997). Existem diversos algoritmos de AM cujo intuito é permitir que, após um determinado treinamento com certo conjunto de dados cujas instâncias têm classificação conhecida, uma máquina seja capaz de interpretar novos dados e classificá-los de maneira apropriada a partir de uma generalização do que lhe foi apresentado anteriormente.

No AM algumas técnicas e algoritmos, como *Naive Bayes* (NB), seção 3.3.1, e SVM (*Support Vector Machines*), seção 3.3.2, são utilizados para classificar um texto. Nesse caso, o sistema, além de aprender a importância de uma palavra-chave óbvia, considera outras palavras que podem ser fundamentais, além da pontuação e da frequência.

O inconveniente do aprendizado de máquina supervisionado, no contexto de análise de sentimentos, é que deve haver vários exemplos de textos já classificados para formar um corpo confiável de treinamento (SILVA, LIMA, BARROS, 2012; QIU et al., 2009). Portanto, muitas vezes é necessário construir exemplos manualmente para auxiliar na detecção mais precisa dos sentimentos e polaridade.

Nas seções 3.3.1 e 3.3.2 serão mostrados algumas técnicas e algoritmos de aprendizagem de máquina mais utilizados para a análise de sentimentos.

3.3.1. *Naive Bayes* (NB)

Segundo Oguri e Milidiú (2006), o *Naive Bayes* é “provavelmente o classificador mais utilizado em *Machine Learning*” (p. 25). É denominado ingênuo (*naive*) por assumir que os atributos são condicionalmente independentes. Em outras palavras, considera-se que as entradas são independentes entre si, o que não ocorre na maioria dos problemas práticos.

Mesmo que esta premissa de ingenuidade seja mantida, o classificador reporta resultados que não comprometem a qualidade.

O *Naive Bayes* é um simples classificador probabilístico baseado na aplicação do teorema de *Bayes*. Ele é frequentemente utilizado como base na classificação de textos por ser rápido e fácil de implementar (RENNIE et al, 2003, p. 1, tradução livre). Gomes (2013) cita que o classificador *Naive Bayes* é considerado um dos mais eficientes em questões relacionadas com processamento e precisão na classificação de novas amostras.

O NB pode ser definido na equação abaixo:

$$P(C_i|\vec{d}) = P(C_i)\frac{P(\vec{d} \mid C_i)}{P(\vec{d})}$$

Equação 1 - Classificador probabilístico baseado no teorema de *Bayes*

Esse tipo de classificador computa a probabilidade de um documento $\vec{d}$ pertencer à classe C_i, assumindo que a presença de um termo em uma categoria não está condicionada a presença de qualquer outro. Devido à independência dos termos, apenas as variações para cada classe necessitam de ser determinada, e não a matriz de covariância completa (ZHANG, 2004). Segundo Domingos e Pazzani (1997), a independência de termos na maioria dos casos não prejudica a eficiência do classificador.

3.3.2. *Support Vector Machines* (SVM)

As Máquinas de Vetores de Suporte (SVMs, do Inglês *Support Vector Machines*) constituem uma técnica de aprendizado que vem recebendo crescente atenção da comunidade de AM (MITCHELL, 1997). Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as Redes Neurais Artificiais (RNAs) (BRAGA, CARVALHO, LUDERMIR, 2000; HAYKIN, 1999). Exemplos de aplicações de sucesso podem ser encontrados em diversos domínios, como na categorização de textos (JOACHIMS, 2002), na análise de imagens (KIM et al., 2002; PONTIL, VERRI, 1998) e em Bioinformática (NOBLE, 2004; SCHÖLKOPF, GUYON, WESTON, 2003).

O SVM é uma técnica recente (da década de 90) utilizada no processo de reconhecimento de padrões e regressão linear. Segundo Duarte e Milidiú (2007), o SVM é uma

técnica de aprendizado linear baseada no uso de *kernels* (núcleo) e de regras não-lineares. A ideia central do SVM é o núcleo do produto interno entre o chamado vetor de suporte e um vetor retirado do espaço de entrada (HAYKIN, 1999). Os vetores de suporte são um subconjunto dos dados de treinamento.

Conforme Burges (1998), *Support Vector Machines* é uma ferramenta que fornece uma nova abordagem para o problema de reconhecimento de padrões. SVM é um classificador supervisionado que tem por objetivo encontrar um hiperplano que separa um conjunto de dados em classes discretas, utilizando-se de processo iterativo e de exemplos de treinamento para ajustar este hiperplano. Para isto, SVM encontra um hiperplano que otimiza a separação das classes, também conhecido como hiperplano ótimo ou ideal, que maximiza a distância entre as classes, sendo usado como fronteira de decisão. (ZHU, BLUMBERG, 2002 apud MOUNTRAKIS et al., 2011; KAVZOGLU, COLKESEN, 2009; PETROPOULOS et al., 2012; COSTA et al., 2010; ANDREOLA, HAERTEL, 2009)

3.4. PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento da Linguagem Natural (PLN) é um conjunto de técnicas computacionais para analisar e representar ocorrências naturais de texto em um ou mais níveis de análise linguística com o objetivo de se alcançar um processamento de linguagem similar ao humano para uma série de tarefas ou aplicações. (LIDDY, 2001, p.1).

Segundo a Comissão Especial de Processamento de Linguagem Natural da Sociedade Brasileira de Computação – SBC[15]:

> "A área de PLN, também denominada Linguística Computacional ou, ainda, Processamento de Línguas Naturais, lida com problemas relacionados à automação da interpretação e da geração da língua humana em aplicações como Tradução Automática, Sumarização Automática de Textos, Ferramentas de Auxílio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação, entre muitas outras...".

Já para Perna, Delgado e Finatto (2010) PLN é uma área de Ciência da Computação que estuda como os computadores podem analisar e/ou gerar textos em linguagem natural.

[15] Disponível em: http://www.nilc.icmc.usp.br/cepln/

Enquanto Lopes (2002) cita que o PLN não é uma tarefa trivial devido à natureza ambígua da linguagem natural. Essa diversidade faz com que o PLN difere do processamento das linguagens de programação de computador, as quais são fortemente definidas para evitar a ambiguidade.

Ferneda (2013) explica que o PLN não se caracteriza como um modelo de recuperação de informação, na medida em que não propõe uma estrutura para a representação dos documentos e não formaliza explicitamente uma função de busca. Porém, é através do PLN que a Recuperação de Informação se aproxima do arsenal metodológico da Inteligência Artificial e viabiliza soluções para alguns de seus problemas.

3.5. O PROCESSO DE MINERAÇÃO DE TEXTO

A mineração de textos pode ser dividida em várias fazes, gerando um processo estruturado onde passos devem ser completados antes de se prosseguir para as próximas fases.

Aranha et al. (2007), Carrilho (2007) propõem um modelo dividido em cinco fases para o processo de MT. Para estes autores, as fases do KDT, são:

1. Coleta
2. Pré-processamento
3. Indexação
4. Mineração ou processamento
5. Pós-processamento ou análise da informação

Figura 5 - Processo de KDT

Fonte: Adaptado de Aranha et al. (2007)

3.5.1. Coleta

Primeiramente, tem-se a etapa de coleta de dados, onde textos são obtidos de diversas fontes através de *web crawlers*, programas que visitam *sites* e extraem informações, para constituir o material a ser analisado no estudo em questão.

Coletar dados é uma atividade trabalhosa. Um dos motivos é que os dados podem não estar disponíveis em um formato apropriado para serem utilizados no processo de mineração de textos.

3.5.2. Pré-processamento

Uma vez que os textos tenham sido coletados e armazenados em uma base de dados, geralmente não estão em formato adequado para extração de conhecimento, faz-se necessária a aplicação de métodos para extração e integração, transformação, limpeza, seleção e redução de volume destes dados, antes da etapa de mineração.

A etapa de pré-processamento é responsável por transformar uma coleção de documentos em uma representação estruturada adequada, normalmente no formato de uma tabela atributo-valor, Tabela 2, a qual é mais apropriada para processamento do que simples arquivos textos. A etapa de pré-processamento geralmente possui um elevado custo computacional, e um organizado e cuidadoso planejamento nesta etapa é fundamental para obter um bom desempenho no processo de mineração de textos (WEISS et al., 2005).

Tabela 2 - Representação atributo/valor

	Atrib1	**...**	**AtribN**
Doc1	V11	...	V1N
...	...	...	...
DocM	VM1	...	VMN

Fonte: Carrilho Junior e Passos (2007)

Existem na literatura duas técnicas muito utilizadas no pré-processamento de textos que serão descritas a seguir.

Uma delas é o *stemming*, este processo reduz termos de diferentes classes gramaticais que possuem mesmo radical para a forma denominada *stem*, sendo eliminados sufixos

provenientes de flexão ou derivação. Por exemplo, a radicalização das palavras *livrinho, livros* e *livro* é *livr*. (GONZALES et al., 2006).

Outra técnica muito utilizada é o *stopwords* que visa à remoção de termos como preposições, artigos e conjunções (GONZALEZ et al., 2006, GONZALEZ; LIMA, 2003) de documentos textuais.

Para a tarefa de análise de sentimentos é importante que quaisquer informações que levam a um entendimento de uma frase estejam bem estruturadas. Por exemplo, a frase "**Eu não gosto do partido, e também não votaria novamente nesse governante!**".

Está claro que a palavra "não" está negando que gosta do partido político e que não voltaria a votar nesse governante, portanto a frase é claramente negativa.

Entretanto, aplicando técnicas de *stopwords* palavras que são muito utilizadas podem ser removidas da frase, distorcendo completamente o sentido das mesmas. Aplicando essa técnica a frase analisada anteriormente poderia ficar da seguinte forma:

"**Gosto partido votaria novamente nesse governante**".

A sentença ganhou um novo sentido, o advérbio de negação "**não**" foi removido mudando completamente o sentido da frase.

Apesar de amplamente utilizada em análise de sentimento e classificação de textos, a remoção das *stopwords* não contribui para a melhoria no desempenho dos classificadores, independentemente do idioma (RIBEIRO, 2015).

Por esse motivo, nesse trabalho não iremos remover *stopwords* e nem fazer *stemming*, pois essas técnicas podem afetar negativamente o desempenho do classificador.

3.5.3. Indexação

A etapa de indexação tem como finalidade representar os documentos de uma maneira inteligível para o computador e facilitar o seu acesso futuro, atuando como um índice.

O objetivo principal da indexação e normalização dos textos é facilitar a identificação de similaridade de significado entre suas palavras, considerando as variações morfológicas e problemas de sinonímia (EBECKEN; LOPES; COSTA, 2003).

Este processo tem como resultado a geração de um índice. Esse índice é construído através de um processo de indexação. Indexar, portanto, significa identificar as características de um documento e colocá-las em uma estrutura denominada índice.

3.5.4. Mineração ou Processamento

O processo de mineração deve ser direcionado para o cumprimento dos objetivos da pesquisa. Na prática, envolve a escolha, configuração e execução de um ou mais algoritmos para extração de conhecimento.

3.5.5. Análise

A última etapa do processo de mineração de texto é responsável pela avaliação e interpretação dos padrões extraídos. Esta etapa visa constatar se o objetivo almejado foi alcançado ou se todas ou algumas etapas do processo necessitam ser refeitas. Os padrões descobertos podem ser avaliados pelo usuário final, especialistas do domínio e analistas de dados, com o intuito de validar o conhecimento obtido (EBECKEN; LOPES; COSTA, 2003).

Normalmente, os resultados obtidos pelos algoritmos de extração da informação são, estatisticamente, avaliados em termos de métricas como: acurácia, precisão, revocação, medida F1 e macro-F1. Essas medidas serão abordadas na Seção 4.5.

4. DESCRIÇÃO DOS EXPERIMENTOS

Nessa seção serão realizados dois experimentos para testar a capacidade de predição do modelo escolhido. O contexto escolhido para ambos os experimentos foi o da política, pois o mesmo está em constante evidência por conta das inúmeras denúncias de corrupção.

O primeiro experimento visa fazer uma análise nos *tweets* relacionados às duas denúncias apresentadas pelo então procurador-geral da república, Rodrigo Janot, contra o atual presidente da república, Michel Temer. A primeira apresentada em 29 de junho de 2017 por corrupção passiva e a segunda apresentada em meados de setembro do mesmo ano pelos crimes de obstrução de justiça e organização criminosa. Nas duas ocasiões as denúncias foram arquivadas pelo plenário da Câmara dos Deputados, dessa forma, impedindo que o Supremo Tribunal Federal (STF) julgasse as denúncias, sendo assim, as denúncias contra Temer só poderão ser eventualmente analisadas após o peemedebista deixar o cargo. Na primeira denúncia, o parecer do deputado Paulo Abi-Ackel (PSDB-MG), que recomendava o arquivamento da acusação formal feita pela procuradoria-geral da república, recebeu 263 votos favoráveis e 277 contrários. Foram, ainda, registradas 19 ausências e 2 abstenções. Já na segunda denúncia, o atual Presidente recebeu o apoio de 251 deputados e 233 votaram contra o peemedebista, ou seja, a favor do prosseguimento da denúncia para STF.

O segundo experimento faz uma análise nos *tweets* relacionados ao julgamento em segunda instância do ex-presidente Luiz Lula que ocorreu em 24 de janeiro de 2018 por corrupção passiva e lavagem de dinheiro. Na ocasião o ex-presidente Lula teve sua condenação mantida pela 8ª Turma do Tribunal Regional Federal da 4ª Região (TRF-4), que era integrada por: Leandro Paulsen, presidente da turma, Victor Luiz dos Santos Laus e João Pedro Gebran Neto, relator do caso. Os três desembargadores votaram e seguiram a decisão do juiz Sérgio Moro, que já havia condenado Lula em primeira instância. Os magistrados ampliaram a pena do ex-presidente de 9 anos e seis meses para 12 anos e 1 mês e autorizam a sua prisão após esgotados todos os recursos na 2ª instância.

Nesse período muitas mensagens ligadas a esses acontecimentos foram publicadas nas redes sociais. Essas mensagens contêm informações relacionadas ao sentimento e opinião da população acerca dos fatos.

O objetivo deste experimento é analisar a polaridade (i. e. positiva, negativa ou neutra) expressa nos *tweets* relacionados às denúncias e ao julgamento, bem como verificar a eficiência

do recurso léxico escolhido para servir como base no desenvolvimento do sistema de apoio a tomada de decisão.

Essa etapa do trabalho se justifica porque é necessário entender o uso das mídias sociais como formas aceitáveis para os intercâmbios entre governo e seus diversos públicos, e que podem, potencialmente, fazer a diferença nas percepções e sentimentos dos cidadãos em relação ao governo (MERGEL, 2013).

4.1. DICIONÁRIO LÉXICO

Classificadores semânticos atribuem sentimentos aos textos usando dicionários, *corpus* e recursos léxicos que contêm palavras com polaridades previamente atribuídas. Os vários sentidos de uma mesma palavra são levados em consideração para fornecer uma classificação mais específica (DUARTE, 2013).

Léxico é o conjunto ou acervo de palavras que um determinado idioma possui. Portanto, a análise léxica estuda as unidades do vocabulário, ou seja, as palavras portadoras de sentido: substantivos, adjetivos, verbos, advérbios entre outras.

Os léxicos são indispensáveis na análise, processamento e geração da língua natural. Segundo Zavaglia (2006), léxicos para serem usados em (PLN) devem possuir informações adequadas e codificadas para que o algoritmo ou programa desenvolvido possa compreendê-lo e executá-lo. Já Trask (2008) conceitua léxico como um acervo de palavras que integram a língua, ou seja, é o vocabulário de uma língua.

Recursos léxicos são as principais ferramentas linguísticas utilizadas na tarefa de análise de sentimentos. Essa técnica é baseada em dicionário léxico de sentimentos, uma espécie de dicionário que ao invés de possuir como conteúdo o significado de cada palavra, possui em seu lugar um significado quantitativo, podendo ser um número variando de -1 a 1, onde -1 é o maior grau para negativo e 1 o maior grau para positivo. Abordagens léxicas assumem que palavras individuais possuem o que é chamado de polaridade prévia, que é uma orientação semântica independente de contexto e que pode ser expressa com um valor numérico ou classe (TABOADA et al. 2011).

4.2. O CLASIFICADOR

O léxico *SentiWordNet* possui uma coleção de termos que são relacionados a valores que indicam o quanto o termo é positivo, negativo ou neutro.

Neste trabalho, a polaridade dos termos extraídos é calculada a partir da informação de positividade e negatividade existente no SWN. Foi feita a extração de todos os termos existentes no arquivo de dados do SWN e em seguida é feito o cálculo do *score* final de cada termo, conforme as Equações*Equação 2* e Equação *3*.

As equações empregadas neste estudo foram baseadas no trabalho de Xu (2010), descreveremos abaixo as equações. Exemplificaremos com a palavra *friendly* como adjetivo, que no SWN possui quatro sentidos com seus respectivos valores positivo e negativo: [0.25, 0.125], [0, 0], [0.375, 0.125], [0.625, 0.25]. Para cada termo e classe gramatical correspondente, subtrai-se o *score* positivo do negativo, como mostra a equação abaixo.

$$Score(term_{sense}) = [posScore] - [negScore]$$

Equação 2 - *Score* do termo

Logo, obtemos $friendly_1$ = (0.25 – 0.125) = 0.125, $friendly_2$ = (0 - 0) = 0, $friendly_3$ = (0.375 – 0.125) = 0.25 e $friendly_4$ = (0.625 – 0.25) = 0.375. Em seguida, é calculado o *score* final pela média aritmética, conforme Equação 3, tendo como base os valores retornados pela Equação Equação *2*.

$$scoreFinal(term) = \frac{1}{n}\sum_{i=1}^{n} score(term)_i$$

Equação 3 - *Score* final do termo

Onde, $score(term)_i$ correspondem aos valores retornados pela Equação 2 e n corresponde a quantidade de *sense* do respectivo termo. A palavra *friendly* possui quatro *sense* e retornou os valores [0.125, 0, 0.25 e 0.375]. O *score* final para o termo será: ((0.125 + 0 + 0.25 + 0.375) / 4), que é igual a ((0.75) / 4), resultando o valor igual a 0.1875.

Para este trabalho, foi considerada classificação negativa, valores abaixo de zero. Classificação positiva, valores acima de zero e para valores iguais a zero, classificação neutra. A palavra *friendly*, para a classe gramatical adjetivo, é classificada como positiva, com um *score* final igual a 0.1875.

4.3. CONSTRUÇÃO DO *CORPUS*

Atualmente, as principais redes sociais *online* provêm interfaces ou serviços para a captura parcial ou total de seus dados. Algumas redes sociais disponibilizam uma API (*Application Programming Interface*) para que estudos ou coletas de dados sejam realizados de maneira simplificada, o que garante que os dados que ali circulam e são coletados estão vinculados a uma conta de desenvolvedor e sob um conjunto de termos de responsabilidade da rede social.

"Os dados do *Twitter* são a fonte mais abrangente de conversas públicas ao vivo em todo o mundo. Nossas APIs REST, *Streaming* e *Enterprise* permitem a análise programática dos *Tweets*" (TWITTER, s.d.).

Basicamente a extração de comentários da rede de informações *Twitter* pode ser realizada de duas maneiras. A primeira forma é a *Twitter search* API[16], uma API disponível no próprio *website* do serviço, a qual disponibiliza os dados no formato JSON (*JavaScript Object Notation*)[17], o que facilita o seu processamento por diversas linguagens, sendo necessária apenas a leitura do arquivo no formato texto.

A segunda forma faz uso da linguagem *Java* (forma utilizada nesse experimento), para qual é disponibilizada uma biblioteca *open source* que realiza o acesso à API do *Twitter*. Segundo Winterwell (2011) essa biblioteca é classificada como robusta e fácil de utilizar. No caso da consulta, uma das vantagens de sua utilização é o fato de seu retorno ser um objeto, característica comum da linguagem *Java*.

Para este trabalho, foram selecionados aproximadamente 1000 *tweets* no contexto da denúncia contra o atual presidente, Michel Temer, utilizando as seguintes estratégias de busca: "denúncia Michel Temer", "Votação da denúncia Michel Temer", "#ficatemer" e "#foratemer". Para o segundo experimento, julgamento em segunda instância do ex-presidente Lula, foram coletados aproximadamente 900 *tweets*, com a seguinte estratégia de busca: "julgamento lula trf4". A coleta foi feita em três momentos diferentes: o primeiro aconteceu nos dias que antecederam o julgamento, para analisar qual era a expectativa da população para o julgamento, e foi chamado de "pré-julgamento". O segundo momento de coleta foi no dia do julgamento, para analisar como seria a reação dos apoiadores e dos movimentos contrários ao ex-presidente

[16] Disponível em: https://dev.twitter.com/docs/api/1/get/search
[17] Disponível em: http://www.json.org/

no momento em que os desembargadores liam as suas sentenças, e foi chamado de "julgamento". O terceiro e último momento de coleta de dados foi um dia após o julgamento para analisar a reação da população ao resultado do julgamento, e foi chamado de "pós julgamento". Os *Tweets* coletados foram salvos em um arquivo CSV [18], para que, posteriormente, pudessem ser analisados.

4.4. PREPARAÇÃO DOS DADOS

Uma vez que os dados estejam disponíveis, ou seja, a etapa de coleta dos *tweets* tenha sido concluída, o próximo passo é realizar o pré-processamento desses dados.

Os *tweets* coletados foram normalizados manualmente a fim de remover referências a usuários, os nomes que vem logo após o @ (arroba), e *urls* que em nada iriam agregar ao sentimento contido na frase, bem como expandir abreviaturas para a frase ficar mais legível.

Para a classificação de sentimentos este trabalho baseou-se na abordagem de DENECKE (2008), a qual utiliza o *SentiWordNet* por meio da tradução do texto original a ser analisado para a língua inglesa, haja vista que o SWN utiliza um dicionário léxico para a língua inglesa e é mais vantajoso e menos trabalhoso traduzir os *tweets* para o inglês que traduzir todo o dicionário para o português, e, então, faz a aplicação do algoritmo para realizar a classificação das palavras do texto. Para a tradução dos textos foi utilizada uma ferramenta *online* (*onlinedoctranslator*)[19], onde por meio dela é possível enviar um arquivo de texto e esse arquivo é traduzido e disponibilizado para *download*.

Tabela 3 - Exemplo de *tweets* tratados

Texto Extraído	Texto tratado para análise	Texto traduzido
O vaivém de ministros escalados para votar a denuncia contra o presidente Michel Temer (PMDB) no Congresso... https://t.co/4G0RV3BqRl;	O vaivém de ministros escalados para votar a denúncia contra o presidente Michel Temer (PMDB) no Congresso	*The swing of ministers scaled to vote the denunciation against President* Michel Temer (PMDB) *in Congress*

[18] CSV: *Comma Separated Values*, é um formato onde há um conjunto de valores em cada linha, separados por vírgula ou outro separador textual como ':' ou ';'.
[19] Disponível em: http://www.onlinedoctranslator.com/pt/

RT @celsopansera: Resultado da condenação de Lula: dólar cai, ricos ganham R$ 100 bilhões na bolsa, especuladores e banqueiros eufóricos	Resultado da condenação de Lula: dólar cai, ricos ganham R$ 100 bilhões na bolsa, especuladores e banqueiros eufóricos	*result of* Lula*'s condemnation: Dollar falls, rich earn* R$ 100 *billion on the stock market, speculators and bankers euphoric*

Fonte: O autor (2018)

Posteriormente os *datasets* normalizados foram submetidos a uma ferramenta desenvolvida pelos pesquisadores da Universidade Federal de Minas Gerais (UFMG) chamada *ifeel*[20], por meio dessa ferramenta é possível submeter um *dataset* e o mesmo ser retornado rotulado, ou seja, nós enviamos o arquivo e ele volta com todas as frases classificadas em positivo, negativo ou neutro. Dessa forma, ao submetermos o *dataset* na nossa aplicação podemos comparar os resultados e assim obter os dados necessários para as avaliações.

Tabela 4 - Exemplos de textos classificados pelo *ifeel*

Texto enviado ao *ifeel*	**Polaridade**
After following closely in the Chamber of Deputies to vote the denunciation of the Attorney General's Office against President Michel Temer.	Positiva
There is no longer any possibility of victory in Janot's *denunciation Chamber of Deputies has just filed the second complaint against.*	Negativa
The former president spoke in the center of São Paulo*, after his sentence extended on trial in* TRF-4*.*	Neutra

Fonte: O autor (2018)

4.5. MÉTRICAS DE AVALIAÇÃO

Para Witten e Frank (2011), métodos estatísticos podem ser utilizados como forma de avaliação dos algoritmos, saber se o processo funcionou ou não como previsto anteriormente. Nesse caso, as métricas podem informar para o usuário percentual de classificações corretas para um determinado contexto.

Segundo Ribeiro et al. (2015) um aspecto chave na avaliação das abordagens para a análise de sentimentos diz respeito às métricas utilizadas. Neste contexto, três métricas

[20] Disponível em: http://www.ifeel.dcc.ufmg.br

principais são comumente empregados para validar a eficiência de um método: acurácia, precisão e revocação.

Ainda para Ribeiro et al. (2015):

> A acurácia indica o percentual de sentenças corretamente classificadas, isto é, a soma de acertos de todas as classes dividida pelo número total de sentenças classificadas [...]. Já a precisão é calculada para cada classe individualmente e evidencia o percentual de sentenças corretamente classificadas para aquela classe. Ou seja, basta dividir os acertos da classe pela quantidade de elementos classificados como pertencendo àquela classe [...]. Enquanto a revocação é calculada justamente pelo total de sentenças corretamente classificadas para uma classe sobre o total de sentenças desta classe na base de dados.

Para melhor representação das métricas utilizadas nesse trabalho, é realizada a matriz confusão representada na Tabela 5. As colunas representam classificações realizadas por especialistas, no caso desse trabalho as classificações foram feitas utilizando a ferramenta *ifeel* que utiliza 18 métodos para classificar a polaridade de uma sentença, e as linhas indicam as classificações geradas pelo método estudado.

Tabela 5 - Matriz Confusão

	Positiva Humana	**Negativa Humana**	**Neutra Humana**
Positiva Máquina	TP	FPneg	FPneu
Negativa Máquina	FNpos	TN	FNneu
Neutra Máquina	Fnpos	Fnneg	Tn

Fonte: O autor (2017)

Para cada classe rotulada manualmente como “Positiva”, são separadas as instâncias corretamente classificadas pelo método (TP – *True Positive*), bem como as classificadas como “Negativa” (FNpos – *False Negative - Positive*) e como “Neutra” (Fnpos – *False Neutral – Positive*). O mesmo se repete nas outras colunas para as classes “Negativa” (FPneg – *False Positive - Negative*), (TN – *True Negative*), (Fnneg – *False Neutral – Negative*) e “Neutra” (FPneu – *False Positive – Neutral*), (FNneu – *False Negative – Neutral*) e (Tn – *True Neutral*).

Dessa forma, a fórmula para se calcular a acurácia é a seguinte.

$$A = \frac{TP + TN + Tn}{TP + FP + TN + Fn + Tn + Fn}$$

Equação 4 - Cálculo da Acurácia

Para se calcular a precisão da classe Positiva (para as outras classes a fórmula é a mesma, bastando apenas substituir os valores) pode-se utilizar a seguinte fórmula.

$$P = \frac{TP}{TP + FPneg + Fpneu}$$

Equação 5 - Cálculo da Precisão

Seguindo o mesmo raciocínio mostrado anteriormente, obtém-se a fórmula para calcular a revocação.

$$R = \frac{TP}{TP + FNpos + Fnpos}$$

Equação 6 - Cálculo da Revocação

E por último, temos a medida F1, que é a média harmônica entre precisão (P) e revocação (R), esta medida é importante para avaliar o desempenho dos classificadores em medida única.

$$F1(classe) = \frac{2 * P(classe) * R(classe)}{P(classe) + R(classe)}$$

Equação 7 - Cálculo da medida F1

A medida F1 final é dada pela macro-F1 (SEBASTIANI, 2002), utilizada para medir a efetividade global de classificação já que a medida F1 se aplica a cada classe individualmente.

A Macro-F1 é calculada com base na média das medidas F1 de cada classe separadamente, independentemente do tamanho relativo de cada classe. Desta forma, a acurácia e a Macro-F1 fornecem parâmetros complementares para a verificação da efetividade de classificação de um método (RIBEIRO et al. 2015).

5. RESULTADOS E DISCUSSÕES

Nessa seção são apresentados e discutidos os resultados obtidos ao aplicar a metodologia proposta mediante a realização dos dois experimentos descritos na seção 4. Na seção 5.1 são apresentados os resultados referentes ao primeiro experimento: denúncia contra o atual presidente da república, Michel Temer. Na seção 5.2 são discutidos os resultados referentes ao segundo experimento: julgamento do ex-presidente Lula.

Para o método estudado, os resultados do processamento da ferramenta foram confrontados com os resultados obtidos pelo método manual e submetidos às medidas de acurácia, precisão, revocação e medida F1, discutidos na seção anterior.

5.1. RESULTADOS DO PRIMEIRO EXPERIMENTO

Na Tabela 6, é mostrado como ficou a matriz confusão preenchida com os valores obtidos do *dataset* referente ao primeiro experimento.

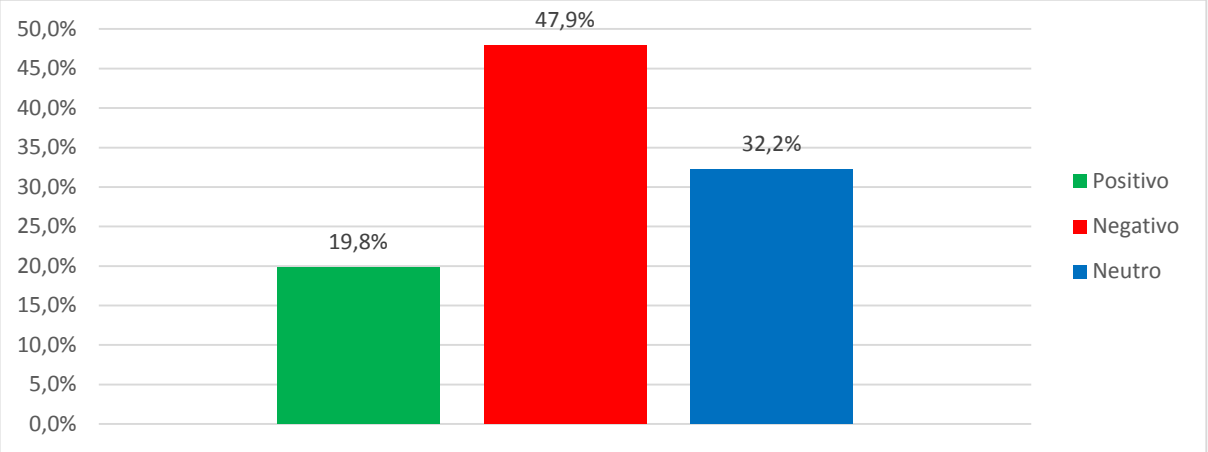
Tabela 6 - Matriz confusão referente ao primeiro experimento

	Positiva Humana	**Negativa Humana**	**Neutra Humana**
Positiva Máquina	16	2	3
Negativa Máquina	6	84	22
Neutra Máquina	4	20	50

Fonte: O autor (2017)

Na Figura 5, é mostrado o percentual de cada sentimento no *dataset* analisado, como já era de se esperar, pelo contexto analisado, a quantidade de *tweets* positivos foi muito baixa, ficando apenas com 19,8% do total, enquanto o sentimento negativo ficou quase com a metade, 47,9% e o sentimento neutro acumulou 32,2% dos comentários.

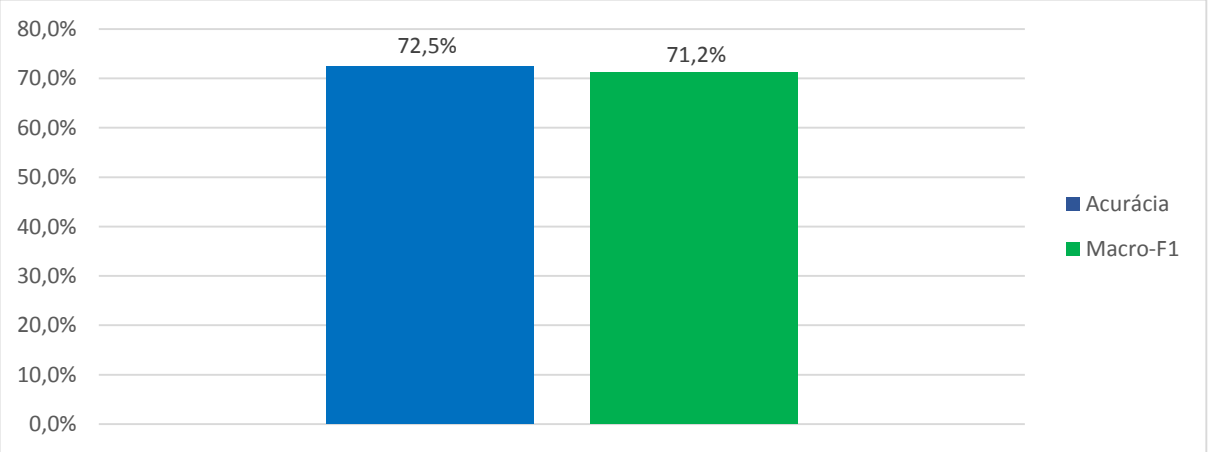
Figura 6 - Percentual de cada sentimento presente no primeiro experimento



Fonte: O autor (2017)

Na Figura 6, abaixo, é mostrada a acurácia e o macro-F1 para o primeiro experimento. Podemos observar que de uma maneira geral o modelo obteve um bom resultado, se levarmos em consideração que uma abordagem não supervisionada tem em média 72% de acurácia e o nosso modelo alcançou uma acurácia de 72,5% enquanto o macro-F1 ficou com 71,2%.

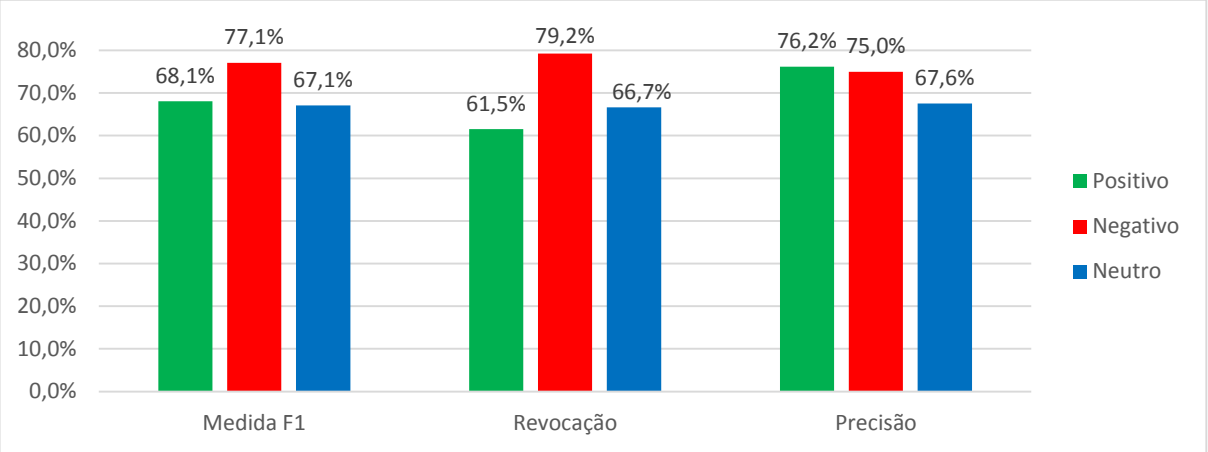
Figura 7 - Acurácia e macro-F1 do primeiro experimento



Fonte: O autor (2017)

Na Figura 7 são mostrados os resultados para precisão, revocação e medida F1 para os sentimentos: positivo, negativo e neutro. Nela podemos observar que o sentimento negativo e o positivo obtiveram uma precisão muito boa para este contexto específico. Por outro lado, o sentimento neutro ficou com uma precisão um pouco abaixo das outras classes.

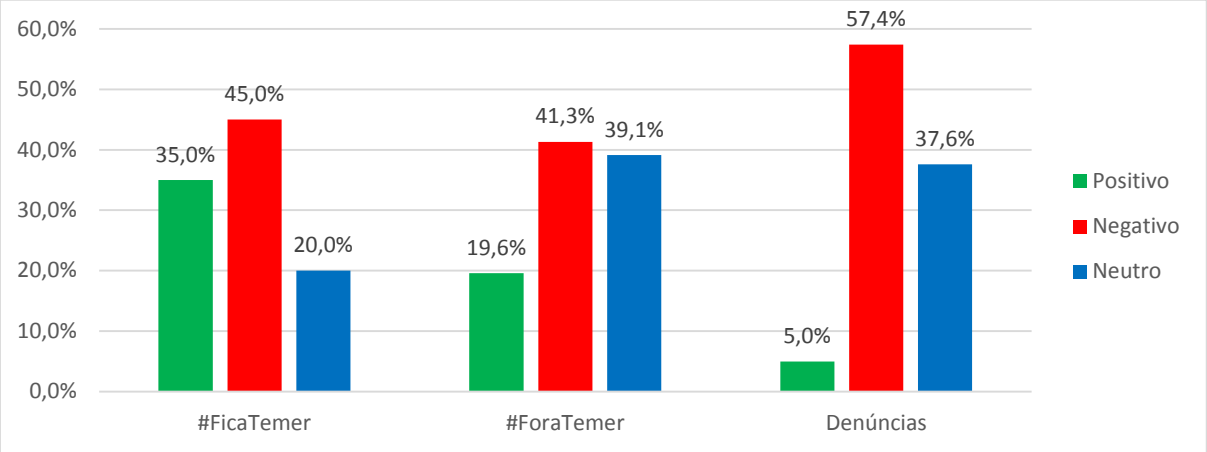
Figura 8 - Precisão, revocação e medida-F1 do primeiro experimento



Fonte: O autor (2017)

Na Figura 8 podemos analisar as polaridades de cada contexto, nela, como era de se esperar, percebemos que a polaridade, quando se refere ao contexto da denúncia, a predominância é de comentários negativos alcançando os 57,4%, enquanto os comentários positivos são de apenas 5%. No contexto #ForaTemer o sentimento negativo domina com 41,3% dos comentários. Agora, quando analisamos o contexto #FicaTemer, onde era esperado que a maioria dos comentários fossem positivos, haja vista que, teoricamente, esses *tweets* pertencessem aos apoiadores do atual presidente, mas o cenário se mostra muito equilibrado, com uma significativa vantagem para os sentimentos negativos (45% dos comentários). Isso se deve ao fato do método que nós estamos estudando, e a maioria dos métodos segue o mesmo caminho, não lidar muito bem com sarcasmos e ironias, como foi abordado na seção 2.4.1. Ao analisar manualmente os *tweets*, podemos perceber a presença muito forte de sarcasmo nos comentários, como pode ser observado a seguir em um comentário de um usuário anônimo do *Twitter*: “#FicaTemer e fode com tudo” (USUÁRIO A., TWITTER, s. d.). Por motivos éticos não serão mencionados os nomes dos autores dos *tweets* mostrados nesse trabalho.

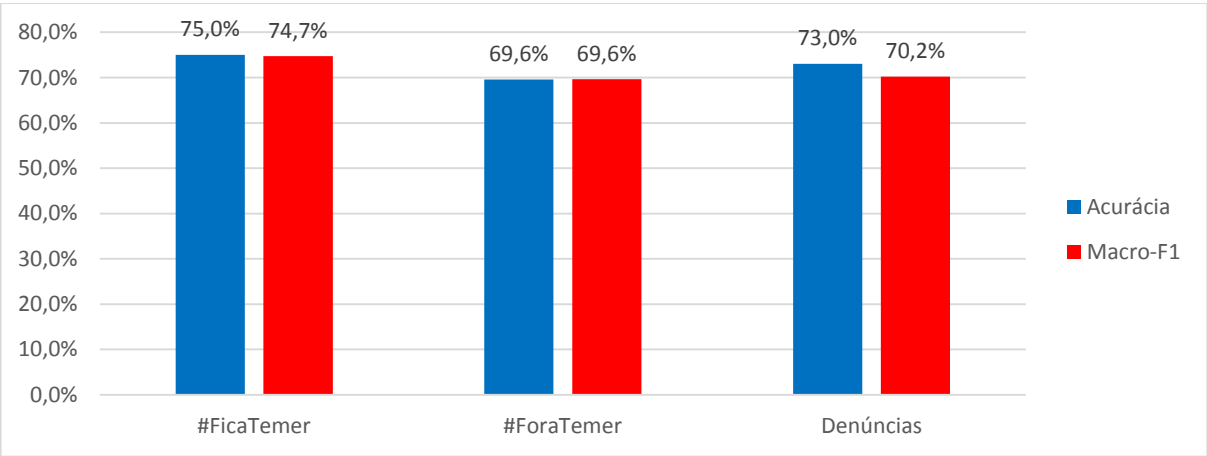
Figura 9 - Polaridade de cada contexto do primeiro experimento



Fonte: O autor (2017)

Na Figura 9 são mostrados os resultados individuais, para cada contexto (#FicaTemer, #ForaTemer e Denúncia), da acurácia e macro-F1. Aqui podemos observar que não houve uma diferença muito grande, em relação a acurácia e a macro-F1, entre os contextos analisados. Portanto, podemos inferir que, nesse caso, o método conseguiu manter uma certa regularidade nas classificações.

Figura 10 - Resultados individuais do primeiro experimento



Fonte: O autor (2017)

5.2. RESULTADOS DO SEGUNDO EXPERIMENTO

Na tabela 7, temos a matriz confusão preenchida a partir do *dataset* referente ao segundo experimento.

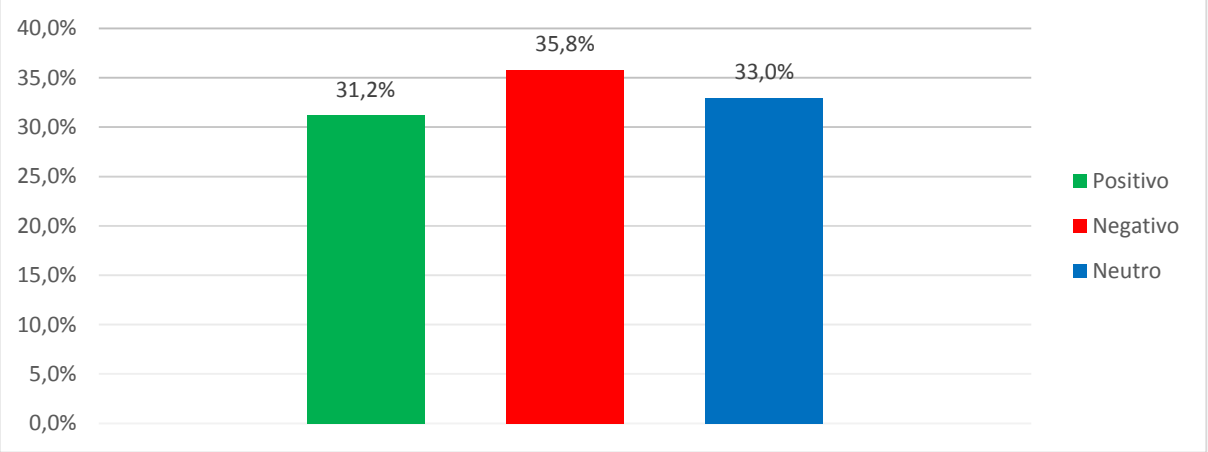
Tabela 7 - Matriz confusão referente ao segundo experimento

	Positiva Humana	Negativa Humana	Neutra Humana
Positiva Máquina	67	7	14
Negativa Máquina	12	82	7
Neutra Máquina	14	14	65

Fonte: O autor (2018)

Na figura 10, abaixo, temos o percentual de cada classe do segundo experimento. Comparativamente, esse cenário obteve um resultado sensivelmente mais equilibrado que o mesmo cenário do experimento anterior encontrado na Figura 5. Isso se deve, muito provavelmente, ao fato de o objeto de estudo desse experimento (Lula) ter uma popularidade e uma quantidade de apoiadores maior que Michel Temer.

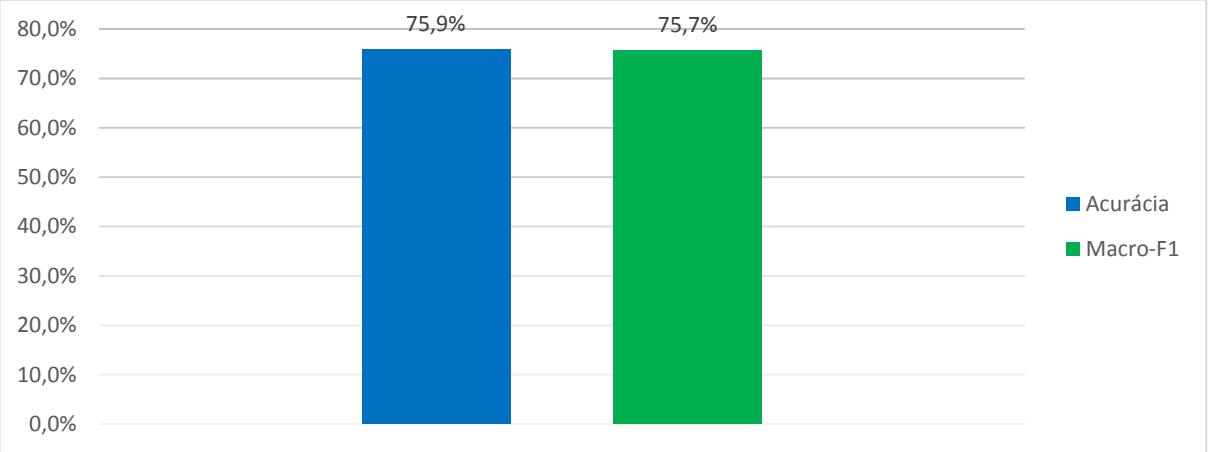
Figura 11 - Percentual de cada sentimento presente no segundo experimento



Fonte: O autor (2018)

Na figura 11, que foca apenas na acurácia e macro-F1 para o segundo experimento, podemos observar que houve uma considerável melhora ao compararmos com o experimento anterior encontrado na Figura 6 que alcançou 63,9% de acurácia e 59,8% de macro F1. Dessa forma, podemos inferir que o fato desse segundo cenário ter um resultado mais equilibrado, no que se refere a polaridade das mensagens, propiciou ao classificador uma melhor capacidade de predição.

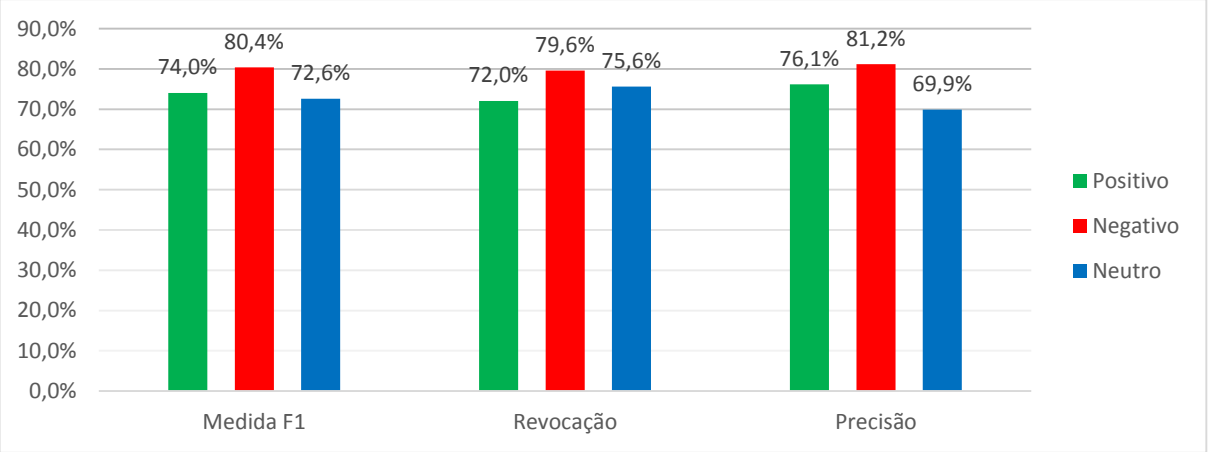
Figura 12 - Acurácia e macro-F1 do segundo experimento



Fonte: O autor (2018)

Na Figura 12 são mostrados os resultados para precisão, revocação e medida F1 para os sentimentos: positivo, negativo e neutro referentes ao segundo experimento.

Figura 13 - Precisão, revocação e medida-F1 do segundo experimento

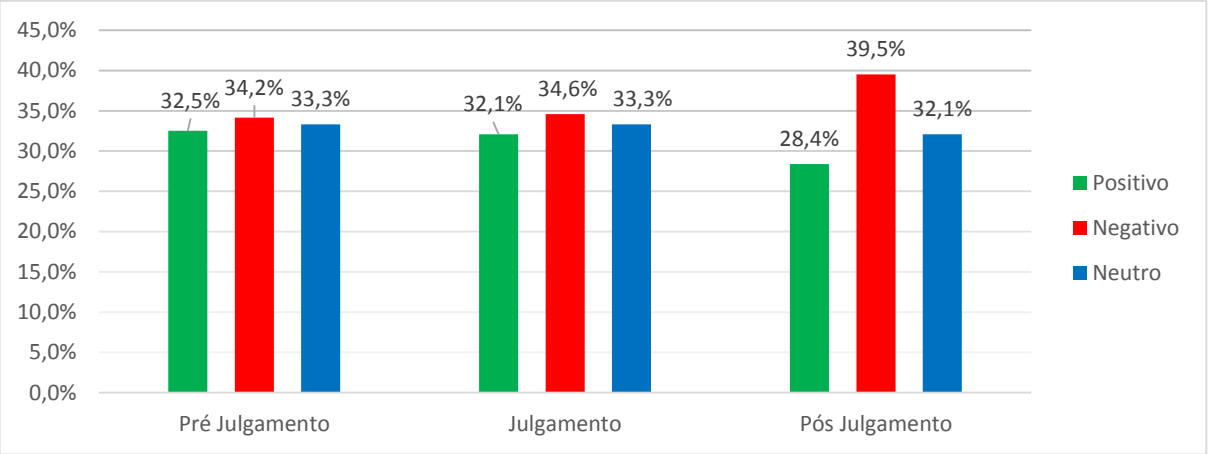


Fonte: O autor (2018)

A Figura 13 apresenta os resultados referentes ao julgamento do ex-presidente Lula. Para essa análise os dados foram separados em momentos diferentes: pré-julgamento, julgamento e pós-julgamento, como foi explicado na seção 4.3. Desse modo, é possível verificar a variação dos sentimentos contidos em cada momento. Vale ressaltar que nos dois primeiros momentos (pré-julgamento e julgamento) houve um certo equilíbrio entre os três sentimentos analisados. No entanto, ao analisar o terceiro cenário (pós-julgamento) pode-se observar que o sentimento negativo obteve uma considerável crescente, tal evento se caracteriza a partir do momento que os desembargadores do TRF-4 decretaram suas sentenças, e dessa forma, mantendo a condenação e aumentando a pena do ex-presidente Lula. Ao tomarem conhecimento do resultado do julgamento, os apoiadores do ex-presidente começaram, como

era de se esperar, a criticar de forma enfática e agressiva a decisão dos desembargadores nas redes sociais. Tal fato pode ser confirmado ao observar o seguinte *tweet:* "Não houve julgamento no TRF-4, mas uma medida de exceção visando a extinção de Lula" (USUÁRIO B., TWITTER, s. d.).

Figura 14 - Polaridade de cada contexto do segundo experimento



Fonte: O autor (2018)

5.3. CONSIDERAÇÕES FINAIS DESSE CAPÍTULO

Esta etapa do trabalho apresentou os resultados obtidos ao aplicar o recurso léxico *SentiWordNet* para o problema de classificação automática de sentimentos no âmbito da política.

Mineração de opinião ou Análise de sentimentos é uma tarefa muito complexa e que ainda não existem ferramentas que a executem com eficácia próxima dos 100%, e para o seu crescimento se faz necessário a criação de novas técnicas e *softwares* que possam auxiliar em pesquisas cada vez mais satisfatórias.

Os resultados obtidos, com os dois experimentos realizados, podem ser considerados satisfatórios, pois segundo (WIEBE et al., 2006) e (GOLDEN, 2011) a capacidade humana de avaliação correta da subjetividade de um texto varia de 72% a 85%, respectivamente, e a média de acurácia dos dois experimentos aqui apresentados foi de 74.2%.

Os resultados apresentados nesse trabalho se mostraram similares aos já obtidos em outros trabalhos nessa área como, por exemplo, em (OHANA; TIERNEY, 2009) e (CAVALCANTI, 2011) que ao utilizarem o *SentiWordNet* em seus experimentos obtiveram uma média de 65,85% e 76% de acurácia, respectivamente, em (KHAN; BASHIR; QAMAR, 2013) que ao realizarem experimentos com dados do *Twitter* obtiveram uma acurácia de 74.2%

e em (OLIVEIRA, 2013) que ao fazer uma análise em comentários retirados do *Twitter*, *YouTube e Mercado Livre* obteve uma acurácia média de 69.4%.

Após a verificação e aprovação dos resultados da análise de sentimentos adotada, foi verificada a polaridade do conjunto de mensagens coletadas durante o período observado. Verificou-se que a maioria das mensagens foi classificada de neutra para negativa. Esse resultado já era esperado por causa do que foi demonstrado por grande parte da população após observar as matérias acerca do assunto publicadas no período analisado.

6. ANTARES

Este trabalho nos propiciou o desenvolvimento de um protótipo de uma ferramenta *web* que nós chamamos de *Antares*. O sistema é capaz de detectar e recuperar dados da *web*, para classifica-los e analisa-los, bem como gerar resumos compreensíveis desses dados.

Nesta seção é feita uma breve descrição do estado atual de desenvolvimento do sistema, passando pela arquitetura geral da aplicação (seção 6.1), as tecnologias que foram utilizadas para o desenvolvimento do protótipo (seção 6.2) e algumas das principais telas do *software* (seção 6.3).

O produto desenvolvido é um *software*, estruturado sobre plataforma *web*, ou seja, será utilizado por meio de aplicações comuns de navegação na *internet* (navegadores). As telas do programa serão exibidas no computador do usuário como qualquer outro *site* de conteúdo disponíveis na *internet.*

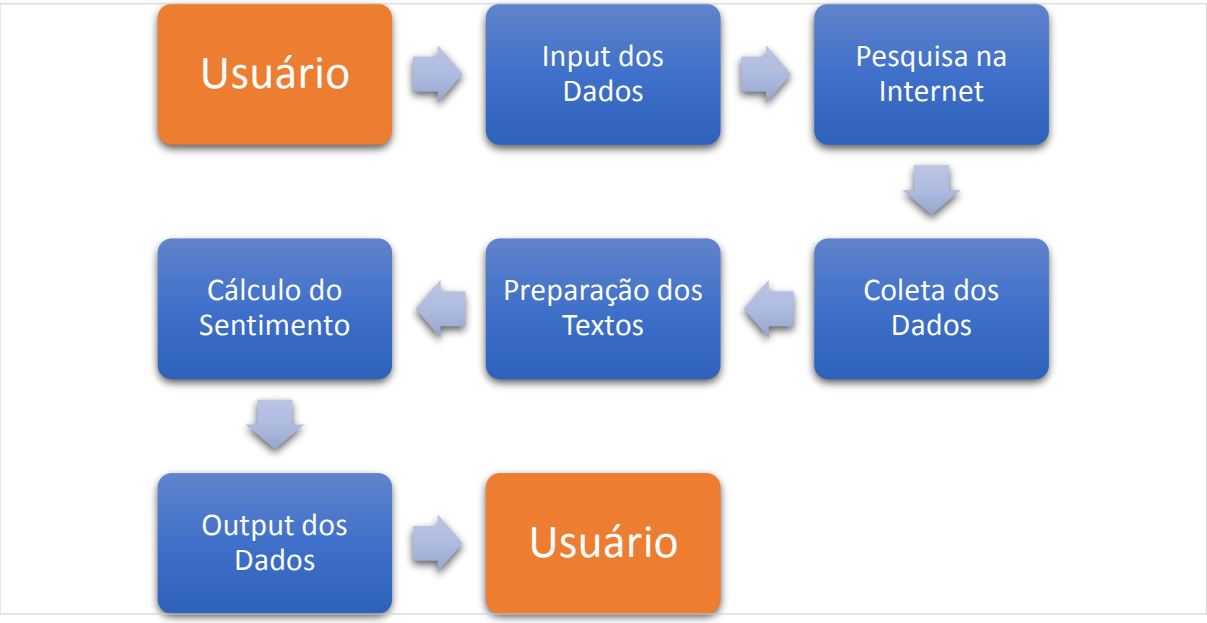
O sistema realiza uma análise probabilística, a partir de uma consulta realizada pelo usuário, em textos públicos retirados de *sites* de notícias, *blogs*, redes sociais entre outros, e cria um resumo classificado e estruturado, dessa forma, facilita o acesso dessas informações para o usuário, já que o sistema retorna o resultado em forma de gráficos e tabelas. Acreditamos que esse sistema será muito útil, pois poderá auxiliar consumidores e empresas a terem uma resposta, em tempo real, sobre produtos e serviços disponíveis.

Através do sistema, o consumidor terá a oportunidade de analisar o que outras pessoas e especialistas que utilizaram aquele serviço ou compraram o produto estão dizendo a respeito, e a partir de então, terão suporte nas tomadas de decisões. Já as empresas contarão com a possibilidade de monitorar as opiniões das pessoas em tempo real a respeito de seus novos produtos ou serviços lançados, bem como avaliar sua própria reputação junto aos consumidores.

6.1. ARQUITETURA GERAL DO SISTEMA

Na Figura 14 é possível acompanhar o fluxo genérico do protótipo desenvolvido até chegar ao resultado final.

Figura 15 – Fluxo do Sistema



Fonte: O autor (2018)

O processo se inicia a partir da entrada do nome de qualquer produto, serviço ou empresa redigida pelo usuário ou pela escolha entre os termos mais pesquisados. A partir de então, o *software* pesquisará pelo termo na imensidão de dados contidos na *internet*.

Para a avaliação dos sentimentos contidos nos textos extraídos é necessário fazer um pré-processamento (seção 4.4), umas das etapas mais importantes de um sistema de análise de sentimentos, pois é quando se prepara todos os dados aplicando técnicas para deixar o texto mais interpretável para o analisador.

Após a realização do pré-processamento, os textos que não estão em inglês são traduzidos, apesar de muitos textos perderem o sentido devido à falta de contexto na hora da tradução. Essa etapa é realizada utilizando o serviço de tradução *online Google Translate*[21]. Embora o serviço não seja gratuito, sendo cobrado pela quantidade de caracteres traduzidos, essa ainda é a técnica mais utilizada por pesquisadores em experimentos relacionados como, por exemplo, em (PORTO et al., 2012; OLIVEIRA, 2013; DUARTE, 2013; REIS et al., 2015; SOUZA; BRANDÃO, 2016), dessa forma, podemos utilizar o método em seu idioma original.

Logo em seguida, o sistema é capaz de verificar a tendência de cada sentimento presente nos textos coletados utilizando o recurso léxico *SentiWordNet* (seção 2.2.1).

[21] Disponível em: http://translate.google.com

Ao fim de todas as etapas, o *software* exibe ao usuário os resultados de todo o processo em forma de gráficos e tabelas com um resumo dos textos encontrados, bem como o sentimento atrelado a cada um deles.

6.2. TECNOLOGIAS ENVOLVIDAS

Por se tratar de uma aplicação *web*, o desenvolvimento do protótipo empregou paradigmas de desenvolvimento para tal finalidade, tais como linguagem de programação e *frameworks* próprios para o desenvolvimento de aplicações *web*. Para o desenvolvimento do protótipo optou-se por utilizar, em sua maioria, ferramentas gratuitas.

Para armazenamento dos dados foi necessária a utilização de um banco de dados. A opção foi pelo *MySQL*[22] sendo gerenciado por um SGBD (Sistema de Gerenciamento de Banco de Dados), que é um dos mais populares, devido à sua otimização para aplicações *web*, disponibilidade para praticamente qualquer sistema operacional e, principalmente, por ser um *software* livre (sob licença GPL - *General Public License*), o que significa que qualquer um pode estudá-lo ou alterá-lo conforme a necessidade.

A codificação das camadas que realizam a busca e coleta na *internet* de todos os dados necessários para a análise, o pré-processamento dos textos capturados e a classificação dos sentimentos, bem como a comunicação entre a aplicação e o banco de dados foi feita em linguagem de programação PHP [23] (um acrônimo recursivo para "PHP: *Hypertext Preprocessor*").

> “O PHP é uma das linguagens mais utilizadas na *web*. Milhões de *sites* no mundo inteiro utilizam PHP. A principal diferença em relação às outras linguagens é a capacidade que o PHP tem de interagir com o mundo *web*, transformando totalmente os *websites* que possuem páginas estáticas” (NIEDERAUER, 2011).

Houve a preocupação também de se desenvolver uma plataforma de fácil manuseio por parte do usuário. Esse requisito foi atingido através do emprego da biblioteca *Twitter Bootstrap*[24], que contém estruturas de *layout* de páginas *web*, folhas de estilo (CSS) e códigos *JavaScript* pré-definidos e que tornam as aplicações *web* mais amigáveis aos usuários,

[22] Disponível em: http://www.mysql.com
[23] Disponível em: http://www.php.net
[24] Disponível em: http://getbootstrap.com/

responsivas e alinhadas com a filosofia *mobile first*. O *layout* das páginas da plataforma foi derivado dos exemplos presentes na documentação deste *framework*.

A arquitetura é distribuída em pacotes que se subdivide por finalidade de utilização das classes, visando atender uma fácil criação, manutenção e separação das camadas do sistema. O padrão adotado para esta divisão foi o MVC[25] (REESE, 2003).

O padrão MVC separa uma aplicação em três camadas distintas e modulares: O modelo (*model*) que normalmente representa os dados por trás da aplicação; A visão (*view*) corresponde o que usuário vê; O controlador (*controller*) que representa a lógica do negócio e a camada de apresentação. Dessa forma, a lógica da aplicação fica isolada da interface do usuário, permitindo ao desenvolvedor alterar, editar e testar cada parte do sistema separadamente. Para a tarefa foi utilizado o *AngularJS*[26].

O *AngularJS* é um *framework* para desenvolvimento *web*, que utiliza a linguagem de programação *JavaScript*. Este *framework* permite estruturação de camadas, criação de componentes modulados e é excelente para o reuso de código, tendo toda uma infraestrutura preparada para integração com servidores *backend*. O *framework* também traz maior facilidade em realização de testes para o desenvolvimento da *web*, evitando possíveis erros (BRANAS, 2014).

Para a apresentação dos resultados em forma de gráfico foi utilizada a biblioteca *ChartJS*[27]. *ChartJS* é uma poderosa biblioteca *JavaScript*, sem dependências, que auxilia na criação de gráficos para *web* através do elemento HTML *canvas*. E o melhor, a utilização do *ChartJS* é uma tarefa extremamente simples até mesmo para *designers* (CUTRELL, 2013).

6.3. PRINCIPAIS TELAS DO SISTEMA

A Figura 15, mostra a tela principal do sistema. Nessa tela é possível o usuário digitar o nome do produto, serviço ou empresa do qual deseja fazer uma análise, a partir de então o usuário é redirecionado para a tela onde é feito um resumo da análise (Figura 17).

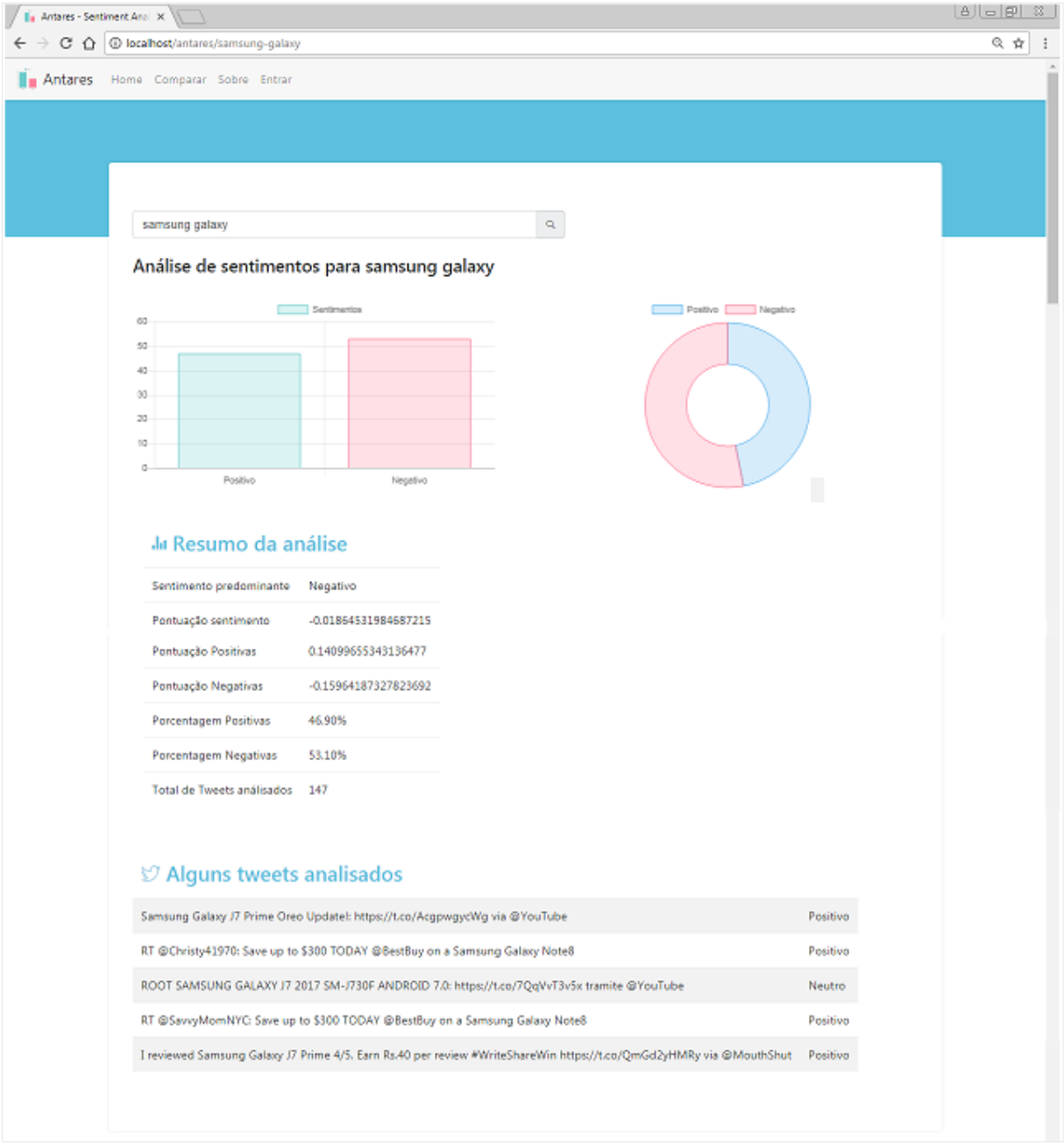
[25] Termo adotado para o padrão de projeto *Model-view-controller* (MVC) (REESE, 2003)
[26] Disponível em: http://angularjs.org
[27] Disponível em: http://www.chartjs.org

Figura 16 - Tela principal

Fonte: O autor (2018)

A figura 16, abaixo, ilustra como são apresentados os resultados obtidos. A interface dispõe de vários tipos de sumarização dos dados analisados tais como gráficos e tabelas que detalham os cálculos realizados e apresentam alguns *tweets* analisados, bem como o sentimento predominante sobre o objeto pesquisado.

Figura 17 - Tela resumo da análise



Fonte: O autor (2018)

Os recursos citados anteriormente podem ser utilizados por qualquer usuário sem ter a necessidade de se cadastrar no *site*. Entretanto, o sistema também permite que o usuário se cadastre a fim de usufruir de outros recursos.

Como o objetivo principal desta monografia não é o desenvolvimento do sistema, outras telas não serão mostradas.

7. CONCLUSÃO

O estudo descrito no presente documento teve como um dos objetivos estudar um classificador não supervisionado desenvolvido para a língua inglesa que faz uso do recurso léxico *SentiWordNet* para o problema de classificação de sentimento. Com base na pesquisa realizada é possível afirmar que o objeto estudado está em constante evolução e se configura como uma solução alternativa e diferenciada para a tomada de decisão.

Inicialmente, foi realizado um levantamento bibliográfico para verificar qual era o estado da arte no tema ao qual estávamos fazendo nossas considerações. Nessa monografia, ainda, foram descritos alguns dos desafios que dificultam a tarefa de realizar a análise de sentimentos, bem como aspectos de alguns recursos léxicos, e neste trabalho foi adotado o recurso léxico *SentiWordNet,* o qual se destacou por ser gratuito tanto para pesquisa acadêmica quanto para sistemas comerciais, e, também, por abranger o maior número de palavras.

Este trabalho mostrou como o processo de Mineração de Textos foi usado para coletar, estruturar o texto extraído e como criar um modelo de classificação de sentimento, que permitiu conhecer a opinião dos brasileiros acerca das recentes denúncias de corrupção na política. O classificador foi desenvolvido com o objetivo de avaliar a capacidade de predição do SWN. O algoritmo foi analisado em dois experimentos no contexto da política no qual demonstrou bons resultados e compatíveis com outros experimentos do mesmo tipo.

Esta monografia, nos possibilitou, ainda, a construção de um protótipo para um sistema *web* que apoiará consumidores e empresas no processo de tomada de decisão.

Nas próximas seções, 7.1 e 7.2, serão apresentadas algumas recomendações de pesquisas para trabalhos futuros e as considerações finais respectivamente.

7.1. RECOMENDAÇÕES PARA TRABALHOS FUTUROS

No decorrer dos experimentos realizados nesta monografia foi possível observar algumas questões e dificuldades que podem ser solucionados em trabalhos futuros.

1. Realizar testes com outros métodos de análise de sentimentos para comparar os resultados obtidos entre os métodos.

2. Alguns léxicos possuem uma boa precisão, mas não contêm palavras específicas de domínio. Outros léxicos contêm uma vasta gama de palavras específicas para domínio, mas não são verificados manualmente, o que aumenta o risco de erros no léxico. Uma maneira de melhorar o desempenho do algoritmo de análise de sentimento é combiná-los entre si e com

técnicas de aprendizado de máquina para chegar ao resultado. Com isso, podem ser obtidos resultados melhores.

3. Refazer o estudo proposto em uma base com comentários de tamanho maior. Embora os resultados deste trabalho tenham sido considerados satisfatórios, eles podem ser melhorados se aplicados a uma base mais extensa de dados.

4. Frases irônicas, de sarcasmo e de *SPAM* podem ser analisadas para adicionar mecanismos nas métricas propostas de como classificar tais frases.

5. O sistema desenvolvido neste trabalho ainda não está totalmente finalizado e que poderá ser melhorado com a introdução de novos módulos.

7.2. CONSIDERAÇÕES FINAIS

Através deste trabalho pode-se observar que a detecção de sentimento tem uma ampla variedade de aplicações em sistemas de informação, incluindo classificação de *reviews*, sumarização de *review* e outras aplicações em tempo real. Há provavelmente muitas outras aplicações que não foram discutidas. Percebe-se que as técnicas de classificação de sentimentos são rigorosamente dependentes do domínio ou tópicos ao qual foram submetidos. Através das comparações realizadas com outros trabalhos realizados na área, ficou evidente que nenhum modelo de classificação consiste em performance superior a outros. É percebido também que diferentes tipos de recursos e algoritmos de classificação são combinados em um modelo mais eficiente, sendo assim são capazes de superar as desvantagens individuais e promover a evolução tanto de um quanto do outro, e finalmente melhorar o desempenho de classificação do sentimento. No futuro, mais trabalhos são necessários para alavancar a capacidade de predição dos métodos existentes.

Analise de sentimentos pode ser aplicada para novas aplicações pelo fato das técnicas e algoritmos utilizados crescerem de uma forma muito rápida. No entanto, muitos problemas nesse campo de estudo permanecem sem solução. Um dos principais aspectos desafiadores existentes é o fato da maioria dos modelos de classificação ser em inglês, o que dificulta a análise, haja vista que é necessário fazer a tradução do que se quer classificar para o inglês, caso esteja em outro idioma, lidar com expressões negativas é outro fator que pode ser levado em consideração. Trabalhos futuros poderiam se dedicar nesses desafios.

BIBLIOGRAFIA

ANDREOLA, R.; HAERTEL, V. **Support Vector Machines na Classificação de Imagens Hiperespectrais**. Simpósio Brasileiro de Sensoriamento Remoto (SBSR). Natal: [s.n.]. 2009. p. 6757-6764.

ARANHA, C. N.; VELLASCO, M. M. B. R.; PASSOS, E. P. L. Uma abordagem de pré-processamento automático para mineração de textos em português: Sob o enfoque da inteligência computacional Pontifícia Universidade Católica do Rio de Janeiro-PUC- RIO. [S.l.]: [s.n.], 2007.

ARAÚJO, M.; GONÇALVES, P.; BENEVENUTO, F. **Métodos para Análise de Sentimentos no Twitter**. Uberlândia: [s.n.]. 2013.

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. SentiWordNet 3.0: An Enhanced Lexical Resource for sentiment analisys and Opinion Mining, Pisa, Italy, 2010.

BACCIANELLA, S.; ESULI, A.; SEBASTIANI, F. **SentiWordNet 3.0:** An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation. [S.l.]: [s.n.]. 2010. p. 2200-2204.

BANFIELD, A. **Unspeakable sentences:** Narration and Representation in the Language of Fiction. Routledge and Kegan Paul. [S.l.]: [s.n.]. 1982.

BARBOSA, L.; FENG, J. Robust sentiment detection on twitter from biased and noisy data. **Proc. of Coling**, 2010.

BELKIN, N. J.; CROFT, W. B. Retrieval tecniques. **Annual Review of Information Science and Tecnology**, p. 112-119, 1987.

BENAMARA, F. et al. Sentiment analysis: Adjectivesand adverbs are better than adjectives alone. **ICWSM '07**, 2007. 203–206.

BIFET, A.; FRANK, E. **Sentiment knowledge discovery in twitter streaming data**. Proc. of 13th International Conference on Discovery Science. [S.l.]: [s.n.]. 2010.

BOGDAN, R. C. E. A. Investigação qualitativa em educação: uma introdução à teoria e aos métodos, 1994.

BOIY, E. et al. Automatic sentiment analysis in on-line text. Proceedings of the 11th International Conference on Electronic Publishing (ELPUB'07), Austria, p. 349-360, 2007.

BOIY, R.; MOENS, M. A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval Journal, Springer Netherlands, v. 12, n.5, p. 526-558, 2008.

BOLLEN, J.; MAO, H. Twitter mood as a stock market predictor. **IEEE Computer, 44(10)**, 2011. 91–94.

BRAGA, A.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais:** Teoria e Aplicações. [S.l.]: Editora LTC, 2000.

BRANAS, R. **AngularJS Essentials**. Birmingham: Packt Publishing Ltd, 2014.

BURGES, C. J. C. **A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery**. [S.l.]: [s.n.]. 1998. p. 121–167.

CAMBRIA, E. et al. **SenticNet 4:** A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japão: [s.n.]. 2016. p. 2666–2677.

CARRILHO JUNIOR, J. R.; PASSOS, E. P. L. **Desenvolvimento de uma Metodologia para Mineração de Textos**. Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, p. 98. 2007. Dissertação de Mestrado.

CARRILHO, J. R. Desenvolvimento de uma metodologia para mineração de textos. Pontifícia Universidade Católica do Rio de Janeiro-PUC-RIO. [S.l.]: [s.n.], 2007.

CAVALCANTI, D. C. **Uma abordagem não supervisionada para classificação de opinião usando o recurso léxico SentiWordNet**. Universidade Federal de Pernambuco. Recife, p. 111. 2011.

CHAKRABARTI, S. **Mining the Web:** Discovering knowledge from hypertext data. Elsevier. [S.l.]: [s.n.]. 2002.

CHAOVALIT, P.; ZHOU, L. Movie review mining: a comparison between supervised and unsupervised classification approaches. Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS), 2005.

COPPIN, B. **Inteligência artificial**. LTC. Rio de Janeiro: [s.n.]. 2010.

CUTRELL, J. Criando um dashboard dinâmico com o ChartJS. **webdesign**, 2013. Disponivel em: <https://webdesign.tutsplus.com/pt/tutorials/build-a-dynamic-dashboard-with-chartjs--webdesign-14363>. Acesso em: 28 abr. 2018.

DAVIDOV, D.; TSUR, O.; RAPPOPORT, A. Enhanced sentiment learning using twitter hashtags and smileys. **Proceedings of Coling**, 2010.

DENECKE, K. **Using SentiWordNet for multilingual sentiment analysis**. Data Engineering Workshop. ICDEW 2008. IEEE 24th International Conference on. [S.l.]: [s.n.]. 2008.

DOMINGOS, P.; PAZZANI, M. **Optimality of the Simple Bayesian Classifier Under Zero-one Loss**. Machine Learning. [S.l.]: [s.n.]. 1997. p. 103.

DUARTE, S. E. **Sentiment Analysis on Twitter for the Portuguese Language**. Faculdade de Ciências e Tecnologia Universidade Nova de Lisboa. [S.l.], p. 188. 2013.

EBECKEN, N.; LOPES, M.; COSTA, M. Mineração de Textos. [S.l.]: [s.n.], v. 1, 2003. Cap. 13, p. 337–370.

ESULI, A. Automatic generation of lexical resources for Opinion Mining: Models, algorithms and applications. Ph.D. thesis, Department of Information Engineering, University of Pisa, Italy, 2008.

ESULI, A.; SEBASTIANI, F. SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), Genova, IT, p. 417–422, 2006.

ESULI, A.; SEBASTIANI, F. SENTIWORDNET: A high-coverage lexical resource for opinion mining. Technical Report 2007-TR-02, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 2007.

FELDMAN, R.; DAGAN, I. **Knowledge discovery in textual databases (KDT)**. KNOWLEDGE DISCOVERY AND DATA MINING. [S.l.]: [s.n.]. 1995. p. 112–117.

FERNEDA, E. **Recuperação da informação: análise da contribuição da ciência da computação para a ciência da informação**. São Paulo, p. 147. 2003.

FILHO, V. M. **e-Recommender: Sistema Inteligente de Recomendação para Comércio Eletrônico**. Escola Politécnica de Pernambuco – Universidade de Pernambuco. Recife, p. 58. 2006.

FRANÇA, T. C. D.; OLIVEIRA, J. **Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013**. BraSNAM - III Brasilian Workshop on Social Networks Analysis and Mining. [S.l.]: [s.n.]. 2014. p. 128-139.

FREITAS, L. D.; VIEIRA, R. **Exploring resources for sentiment analisys in portuguese language**. IEEE. 2015 Brazilian Conference on intelegent Systems (BRACIS). [S.l.]: [s.n.]. 2015. p. 152-156.

GARCIA, D.; SCHWEITZER, F. **Emotions in Product Reviews - Empirics and Models**. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational. [S.l.]: [s.n.]. 2011. p. 483-488.

GIBBS, R. W. **On the psycholinguistics of sarcasm**. Journal of Experimental Psychology: General. [S.l.]: [s.n.]. 1986. p. 3.

GIBBS, R. W.; COLSTON, H. L. **Irony in language and thought:** A cognitive science reader. Lawrence Erlbaum. [S.l.]: [s.n.]. 2007.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. **CS224N Project Report, Stanford**, 2009.

GOLDEN, P. **Write here, write now**. Disponivel em: <http://www.research-live.com/features/write-here-write-now/4005303.article>.

GOMEZ, H. J. C. **Text Mining:** análise de sentimentos na classificação de notícias. Information Systems and Technologies (CISTI). 8th Iberian Conference on. Lisboa: [s.n.]. 2013.

GONZALEZ, M.; LIMA, V. L. Recuperação de informação e processamento da linguagem natural, XXIII Congresso da Sociedade Brasileira de Computação, Anais da III Jornada de Mini-Cursos de Inteligência Artificial, v. 3, Campinas, p. 347-395, 2003.

GONZALEZ, M.; LIMA, V. L.; LIMA, J. V. Termos, relacionamentos e representatividade na indexação de texto para recuperação de informação. Letras de Hoje, v. 41, n. 2, Porto Alegre, p. 65-87, 2006.

GONZÁLEZ-IBÁÑEZ, R.; MURESAN, S.; WACHOLDER, N. **Identifying sarcasm in Twitter:** a closer look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers. [S.l.]: [s.n.]. 2011.

GUPTA, V.; GUPTA, G. A survey of text mining techniques and applications. **Journal of Emerging Technologies in Web Intelligence**, p. 60–76, 2009.

HAYKIN, S. **Neural Networks - A Compreensive Foundation**. Prentice-Hall. New Jersey: [s.n.]. 1999.

HAYKIN, S. S. **Neural Networks:** A Comprehensive Foundation. [S.l.]: Prentice-Hall, 1999. 842 p.

HEARST, M. A. **Untangling text data mining**. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Stroudsburg: [s.n.]. 1999. p. 3–10.

HIROSHI, K.; TETSUYA, N.; HIDEO, W. Deeper sentiment analysis using machine translation technology. **COLING '04**, Morristown, NJ, USA, 2004.

HU, M.; LIU., B. Mining and summarizing customer reviews. **KDD-2004**, 2004. 168–177.

INDURKHYA, N.; DAMERAU, F. J. Handbook of natural language processing. 2 ed. Florida: CRC Press, p. 666, 2010.

JANSEN, B. et al. Twitter power: Tweets as electronic word of mouth. **Journal of the American Society for Information Science and Technology 60(11)**, p. 2169–2188, 2009.

JOACHIMS, T. **Learning to classify texts using support vector machines:** methods, theory and algorithms. Kluwer Academic Publishers. [S.l.]: [s.n.]. 2002.

KAMBER, M.; HAN, J. Data mining: concepts and techniques. **Morgan Kaufmann**, 2001.

KHAN, F. H.; BASHIR, S.; QAMAR, U. TOM: Twitter opinion mining framework using hybrid classification scheme. **Elsevier**, Islamabad, p. 245-257, 21 Setembro 2013.

KIM, K. I. et al. **Support vector machines for texture classification**. IEEE Transactions on Pattern Analysis and Machine Intelligence. [S.l.]: [s.n.]. 2002. p. 1542–1550.

KREUZ, R. J.; GINA, M. C. **Lexical influences on the perception of sarcasm**. Proceedings of the Workshop on Computational Approaches to Figurative Language. [S.l.]: [s.n.]. 2007.

KREUZ, R. J.; SAM, G. **How to be sarcastic:** The echoic reminder theory of verbal irony. Journal of Experimental Psychology: General. [S.l.]: [s.n.]. 1989. p. 374.

LE COADIC, Y. F. **A ciência da informação**. Brasília: Briquet de Lemos. 2004.

LI, X. et al. News impact on stock price return via sentiment analysis. **Knowledge-Based Systems 69**, 2014. 14–23.

LIDDY, E. Natural Language Processing. **Encyclopedia of Library and Information Science**, New York, 2001.

LIU, B. Sentiment analisys and subjectity. Handbook of Natural Language Processing, CRC, Taylor and Francis Group, Second Edition, p. 1-38, 2010.

LIU, B. Sentiment Analysis and Opinion Mining. **Morgan & Claypool Publishers**, 2012.

LIWC. Linguistic Inquiry and Word Count. Disponivel em: <https://liwc.wpengine.com/>. Acesso em: Novembro 2017.

LOPES, I. L. **Uso das linguagens controlada e natural em bases de dados:** revisão da literatura. Ciência da informação. [S.l.]: SciELO Brasil. 2002. p. 41–52.

LÓPEZ YEPEZ, J. **El analisecualitativo de citas como instrumento para el estudio de la criación y transmissión de las ideas cientificas**. Documentación de las ciencias de la información. [S.l.]: [s.n.]. 2003. p. 41-70.

MACEDO, H. Discursos de ódio na Internet e como a IA pode ajudar nessa luta. **Saense**, 2017. Disponivel em: <http://www.saense.com.br/2017/06/discursos-de-odio-na-internet-e-como-a-ia-pode-ajudar-nessa-luta/>. Acesso em: Novembro 2017.

MACIAS-CHAPULA, C. A. **O papel da informetria e da cienciometria e sua perspectiva nacional e internacional.** Brasilia: [s.n.]. 1998. p. 134-140.

MEADOWS, A. J. **A comunicação ciêntifica:** Briquet de Lemos. Brasília: [s.n.]. 1999.

MEDHAT, W.; HASSAN, A.; KORASHY, H. **Sentiment analisys algorithms and applications:** A survey. Ain Shams Engineering Journal, Elsevier, v. 5, n. 4. [S.l.]: [s.n.]. 2014. p. 1093-1113.

MIAO, Q.; LI, Q.; DAI, R. AMAZING: system. Expert Systems with Applications: An International Journal, v. 36, n.3, pp. 7198-7198, 2009.

MIHALCEA, R.; BANEA, C.; WIEBE, J. Learning multilingual subjective language via cross-lingual projections. **ACL '07**, 2007. 976–983.

MILLER, G. A. Wordnet: a lexical database for english. Communications of the ACM, p. 38, 1995.

MITCHELL, T. Machine Learning. McGraw Hill, 1997.

MOURA, M. F. Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos. **Embrapa Informática Agropecuária**, 2004.

MUKHERJEE, S. **Sentiment analysis - a literature survey**. Indian Institute of Technology. Bombay: [s.n.]. 2012.

MUKRAS, R. Representation and learning schemes for sentiment analysis. Ph.D. thesis, The Robert Gordon University, Scotland, 2009.

MUNDIAL, B. **Relatório sobre o desenvolvimento mundial 2016. DIVIDENDOS DIGITAIS: visão geral**. Washington, DC. 2016.

NIEDERAUER, J. **Desenvolvendo Websites com PHP**. 2ª. ed. São Paulo: Novatec Editora, 2011. ISBN 978-85-7522-234-8.

NIELSEN, F. A. **A new ANEW:** Evaluation of a word list for sentiment analysis in microblogs. CEUR Workshop Proceedings. [S.l.]: [s.n.]. 2011. p. 93-98.

NOBLE, W. A. **Support vector machine applications in computational biology**. B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, Kernel Methods in computational biology. [S.l.]: MIT Press. 2004. p. 71–92.

O'CONNOR, B. et al. From tweets to polls: Linking text sentiment to public opinion time series. **Proceedings of ICWSM**, 2010.

OGURI, P.; MILIDIÚ, R. L. **Machine Learning Algorithms for Portuguese Named Entity Recognition**. Rio de Janeiro, p. 10. 2007.

OHANA, B.; TIERNEY, B. **Sentiment classification of reviews using SentiWordNet**. 9th. IT&T Conference. Dublin: [s.n.]. 2009.

OLIVEIRA, D. J. S.; BERMEJO, P. H. D. S. Mídias sociais e administração pública: análise do sentimento social perante a atuação do Governo Federal brasileiro. **O&S**, Salvador, v. 24, n. 82, p. 491-508, Jul./Set. 2017.

OLIVEIRA, F. W. C. D. **Análise de sentimentos de comentários em português utilizando o SentiWordNet**. Universidade Estadual de Maringá. Maringá, p. 45. 2013.

OSIEK, B. A. **Reconhecimento de sentimento em texto abordado através da computação afetiva**. Universidade Federal do Rio de Janeiro, COPPE. Rio de Janeiro, p. 175. 2014.

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

PANG, B.; LEE, L. Opinion mining and sentiment analisys. **Foundations and Trends in information Retrieval**, p. 1-135, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. **Thumbs up?:** sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Stroudsburg, PA: USA: Association for Computational Linguistics. 2002. p. 79–86.

PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. **Linguistic Inquiry and Word Count.** Lawerence Erlbaum Associates. Mahwah: [s.n.]. 2001.

PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. **Linguagens Especializadas em Corpora:** modos de dizer e interfaces de pesquisa. EDIPUCRS. [S.l.]: [s.n.]. 2010.

POIRIER, D. et al. Automating opinion analisys in film reviews: The case of statistic versus linguistic approuch. Proceedings of the LREC 2008 Workshop on Sentiment Analisy: Enotion, Metaphor, Ontology and Terminlogy, p. 94-101, 2008.

PONTIL, M.; VERRI, A. **Support vector machines for 3-D object recognition**. IEEE Transactions on Pattern Analysis and Machine Intelligence. [S.l.]: [s.n.]. 1998. p. 637–646.

PORTO, B. R. M. A. et al. **ANÁLISE DE SENTIMENTO SOBRE VEÍCULOS EM REDES SOCIAIS**. XIII Encontro Nacional de Pesquisa em Ciência da Informação. Belo Horizonte: [s.n.]. 2012.

QIU, G. et al. Domain specific opinion retrieval. In: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology. Sapporo, Japan: Springer-Verlag, p. 318–329, 2009.

QUEIROZ, F. M.; NORONHA, D. P. **Temática das dissertações e teses em ciências da informação no programa de pós graduação em Ciências da Comunicação da USP**. Ciência da Informação. Brasília: [s.n.]. 2004. p. 132-142.

REESE, G. Java Database Best Practice. [S.l.]: [s.n.], 2003. p. 16-17.

REIS, J. et al. **Uma abordagem multilíngue para análise de sentimentos**. In CSBC 2015 - BraSNAM. [S.l.]: [s.n.]. 2015.

RENNIE, J. D. M. et al. **Tackling the poor assumptions of naive bayes text classifiers**. ICML. [S.l.]: [s.n.]. 2003. p. 616-623.

RIBEIRO, F.; ARAÚJO, M.; BENEVENUTO, F. Métodos para análise de sentimentos em mídias sociais. Brazilian Symposium on Multimedia and the Web. (Webmedia), 2015.

RIBEIRO, L. B. **Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: Estudo do impacto do pré-processamento**. Universidade Nacional de Brasília UnB. Brasília, p. 82. 2015.

RILL, S. et al. Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. **Knowledge-Based Systems 69**, 2014. 14–23.

SALMI, M. H. S. **Investigação de Algoritmos de Análise de Sentimento para a Língua Portuguesa**. UEG / CCET. Anápolis, p. 47. 2015.

SALTON, G.; MCGILL, M. J. Introdution to modern information retrieval. **Computer Science Series**, USA, 1983.

SANTOS, C. N. D. **Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro**. Rio de Janeiro. 2005.

SANTOS, R. N. M. **Indicadores estratégicos em ciências e tecnologia:** refletindo a sua prática como dispositivo de inclusão. Transformação. [S.l.]: [s.n.]. 2003. p. 129-140.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspec. Ci. Inf.**, Belo Horizonte, 1996. 41-62.

SCHÖLKOPF, B.; GUYON, I.; WESTON, J. **Statistical learning and kernel methods in bioinformatics**. P. Frasconi and R. Shamir, editors, Artificial Intelligence and Heuristic Methods in Bioinformatics. [S.l.]: IOS Press. 2003. p. 1–21.

SEBASTIANI, F. Machine learning in automated text categorization. ACM computing surveys (CSUR), ACM, v. 4, n. 1, p. 1-47, 2002.

SENTISTRENGTH. SentiStrength, Novembro 2017. Disponivel em: <http://sentistrength.wlv.ac.uk/>.

SENTIWORDNET. Disponivel em: <http://sentiwordnet.isti.cnr.it/.>. Acesso em: Novembro 2017.

SILVA, L. L. A. A. **Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo**. UnB. Brasília, p. 129. 2013.

SILVA, N. R.; LIMA, D. SAPair: Um Processo de AnÁlise de Sentimento no NÍvel de CaracterÍstica, 2012.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de análise de sentimento no nível de característica. In: 4nd International Workshop on Web and Text Intelligence (WTI'12), Curitiba, 2012.

SILVA, R. G. N. E. **Sistema de Recomendação baseado em conteúdo textual: avaliação e comparação**. Universidade Estadual de Feira de Santana, Universidade Federal da Bahia, Programa Multi–institucional em Ciência da Computação. Feira de Santana, p. 122. 2014.

SMITH, A. The Internet and Campaign 2010. Disponivel em: <http://www.pewinternet.org/Reports/2011/The-Internet-and-Campaign-2010/Summary.aspx>. Acesso em: Novembro 2017.

SOUZA, C. H. P. D.; BRANDÃO, W. C. Análise Comparativa de Abordagens de Análise de Sentimento Utilizando Tweets em Língua Portuguesa, Belo Horizonte, 2016.

SPINAK, E. **Indicadores Cienciométricos**. Ciência da Informação. Brasília: [s.n.]. 1998. p. 141-148.

STRITESKY, V.; STRANSKA, A.; DRABIK, P. **Crisis communication on Facebook**. Studia commercialia Bratislavensia. [S.l.]: [s.n.]. 2015. p. 103-111.

TABOADA, M. et al. Lexicon-Based Methods for Sentiment Analysis. **Computational Linguistics**, 37, n. 2, 2011. 267–307.

TAN, A.-H. **Text mining:** The state of the art and the challenges. PROCEEDINGS OF THE PAKDD 1999 WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES. Beijing: [s.n.]. 1999.

TEXTMININGNEWS, 2011. Disponivel em: <http://www.textminingnews.com>. Acesso em: Novembro 2017.

THELWALL, M. et al. **Sentiment strength detection in short informal text**. Journal of the American Society for Information Science and Technology. [S.l.]: [s.n.]. 2010. p. 2544-2558.

TSUR, O.; DAVIDOV, D.; RAPPOPORT, A. **A Great Catchy Name:** Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM-2010). [S.l.]: [s.n.]. 2010.

TUMASJAN, A. et al. Predicting elections with twitter: What 140 characters reveal about political sentiment. **Proceedings of ICWSM.**, 2010.

TWITTER. Twitter. Disponivel em: <https://developer.twitter.com/>. Acesso em: Novembro 2017.

USPENSKY, B. **A Poetics of Composition**. Berkeley, California: University of California Press. 1973.

UTSUMI, A. **Verbal irony as implicit display of ironic environment:** Distinguishing ironic utterances from nonirony. Journal of Pragmatics. [S.l.]: [s.n.]. 2000. p. 1777-1806.

WEISS, S. M. N.; INDURKHYA, T.; DAMERAU, F. J. From textual information to numerical vectors. In Text Mining: Predictive Methods for Analysing Unstructured Information, Springer Verlag, p. 15–44, 2005.

WIEBE, J. **Tracking point of view in narrative. Computational Linguistics**. [S.l.]: [s.n.]. 1994.

WIEBE, J.; WILSON, T.; CARDIE, C. Annotating Expressions of Opinions andEmotions in Language. **Language Resources and Evaluation**, 2006. 165 –210.

WINTERWELL, A. JTwitter - the java library for the Twitter API, 2011. Disponivel em: <http://www.winterwell.com/software/jtwitter.php>. Acesso em: Novembro 2017.

WITTEN, I. H.; FRANK, E. Data mining: Pratical machine learning tools and techniques.[S1]:Morgan Kauffman, 2011.

WIVES, L. **Tecnologias de descoberta de conhecimento em textos aaplicadas à inteligência competitiva**. Exame de Qualificação EQ-069. [S.l.]: [s.n.]. 2002.

XU, Z. A Sentiment analysis model integrating multiple algorithms and diverse features. M. S. thesis, Computer Science and Engineering, Ohio State University, 2010.

YAO, J. et al. Using bilingual lexicon to judge sentiment orientation of chinese words, 2006.

YESSENOV, K.; MISAILOV, S. Sentiment analysis of movie review comments., 2009. Disponivel em: <http://people.csail.mit.edu/kuat/courses/6.863/report.pdf>. Acesso em: Outubro 2017.

ZHANG, H. **The optimality of naive bayes**. Proceedings Of The Seventeenth International Florida Artificial Intelligence Research Society Conference. Miami Beach: [s.n.]. 2004. p. 562-567.