**GABRIELA MARIA MACHADO**

# UM ESTUDO SOBRE ANÁLISE DE SENTIMENTOS DE DADOS COLETADOS DO TWITTER

Monografia apresentada ao PECE – Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para conclusão do curso de MBA em Tecnologia de Software.

São Paulo
2018

**GABRIELA MARIA MACHADO**

# UM ESTUDO SOBRE A ANÁLISE DE SENTIMENTOS DE DADOS COLETADOS DO TWITTER

Monografia apresentada ao PECE – Programa de Educação Continuada em Engenharia da Escola Politécnica da Universidade de São Paulo como parte dos requisitos para a conclusão do curso de MBA em Tecnologia de Software.

Área de Concentração: Tecnologia de Software

Orientador: Prof. Dr. Jorge Rady

São Paulo
2018

Catalogação-na-publicação

Machado, Gabriela Maria
UM ESTUDO SOBRE ANÁLISE DE SENTIMENTOS DE DADOS COLETADOS DO TWITTER / G. M. Machado -- São Paulo, 2018.
49 p.

Monografia (MBA em Tecnologia de Software) - Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.BIG DATA 2.MINERAÇÃO DE DADOS 3.TWITTER I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

**DICATÓRIA**

*Dedico este trabalho à minha família e entes queridos.*

## AGRADECIMENTOS

À Universidade de São Paulo – USP que me possibilitou imenso aprendizado ao oferecer este curso.

Ao Professor Dr. Jorge que foi sempre muito prestativo, atencioso e competente na orientação deste trabalho.

Ao Bruno Paulinelli que muito me ajudou na definição do tema e partes técnicas do trabalho.

A XDani que leu, releu e sugeriu correções em diversas partes do texto.

Aos meus pais, irmãos e sobrinho, que sempre torceram e me apoiaram de todas as formas possíveis.

Aos meus amigos que ficaram ao meu lado nos diversos momentos difíceis.

**RESUMO**

Sabendo da importância que empresas e pesquisadores em geral têm dado para os dados oriundos das Redes Sociais Online (RSO), este trabalho detalha, inicialmente, algumas funcionalidades e características do Twitter, conceitos básicos de Big Data, métodos de coleta e armazenamento de dados do Twitter e algumas ferramentas do Ecossistema Hadoop, apresentando técnicas utilizadas na tarefa de ingestão de grandes dados. Em seguida, são apresentados alguns procedimentos de obtenção da polaridade dos sentimentos dos tweets coletados e os principais benefícios de cada um. Por fim, apresentam-se as vantagens e desvantagens de alguns métodos com relação a outros em determinadas situações, evidenciando comparações entre os procedimentos de análise antes e após o pré-processamento, mostrando em números a importância desta tarefa em meio ao processo como um todo.

Palavras-chave: Big Data. Hadoop. Análise de Sentimentos. Twitter.

**ABSTRACT**

Knowing the importance that companies and researchers in general have given to the data coming from Online Social Networks (OSN), this work first details some features and characteristics of Twitter, basic concepts of Big Data, methods of data collection and storage Twitter and some tools of the Hadoop Ecosystem, presenting techniques used in the task of ingesting Big Data. Following are some procedures to obtain the polarity of the sentiment of the tweets collected and the main benefits of each. Finally, we present the advantages and disadvantages of some methods in relation to others in certain situations, evidencing comparisons between the procedures of analysis before and after the preprocessing, showing in numbers the importance of this task in the middle of the process as a whole.

Keywords: Big Data. Hadoop. Sentiment Analysis. Twitter.

# LISTA DE ILUSTRAÇÕES

## LISTA DE QUADROS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| APIs | APPLICATION PROGRAMMING INTERFACE |
| HDFS | SISTEMA DE ARQUIVOS DISTRIBUÍDOS HADOOP |
| IP | INTERNET PROTOCOL |
| LIWC | LINGUISTIC INQUIRY AND WORD COUNT |
| LR | LOGISTIC REGRESSION |
| ME | ENTROPIA MÁXIMA |
| NB | NAIVE BAYES |
| PLN | PROCESSAMENTO DE LINGUAGEM NATURAL |
| RF | RANDOM FOREST |
| RSO | REDES SOCIAIS ONLINE |
| SGBD | SISTEMAS DE GERENCIAMENTO DE BANCO DE DADOS |
| TS | TÉCNICA SUPERVISIONADA |
| TNS | TÉCNICA NÃO SUPERVISIONADA |
| URL | UNIFORM RESOURCE LOCATOR |

## SUMÁRIO

# 1 INTRODUÇÃO

## 1.1 Motivação

As redes sociais têm despertado muito interesse dos usuários da Internet nos últimos anos, dentre elas, destaca-se o Twitter, que é utilizado para postar comentários através de status curtos. Os mais de 500 milhões de tweets postados diariamente, (número obtido em França (2014)), podem ser submetidos à chamada Análise de Sentimento, tornando-se uma excelente fonte para reunir opiniões de consumidores. Os usuários podem avaliar a reputação das marcas e a qualidade das empresas e estas têm tido o interesse de detectar e analisar tais dados digitais, a fim de melhorar a sua reputação entre os consumidores. Os dados dessas redes podem ser usados para uma série de propósitos, como previsões em geral, marketing, análise de sentimentos, etc. As empresas podem melhorar sua competitividade atendendo às necessidades e expectativas dos consumidores, detectando o sentimento de um tweet e como ele influencia as pessoas, quando a popularidade do autor é levada em conta. A análise de sentimento torna possível identificar as opiniões positivas, negativas ou neutras com base no texto em um determinado tópico.

Entretanto, lidar com grandes quantidades de dados não estruturados não é uma tarefa simples. Uma das formas mais adotadas pelas grandes empresas para análise e armazenamento inteligente de dados não estruturados utiliza o Ecossistema Hadoop, já que dados baseados em texto não se encaixam naturalmente em bancos de dados relacionais.

## 1.2 Objetivo

Sabida a importância das opiniões postadas pelos internautas nas RSO, tem-se como um dos objetivos deste trabalho detalhar métodos de coleta e armazenamento de dados do Twitter, introduzindo o conceito de Big Data e algumas ferramentas do Ecossistema Hadoop, que contempla hoje as principais técnicas utilizadas nesta tarefa de ingestão de grandes dados. Está no escopo deste trabalho exemplificar como isto pode ser feito e quais as maiores dificuldades e impedimentos técnicos e burocráticos envolvidos.

Uma vez disponibilizados os dados, pretende-se apresentar os procedimentos de obtenção da polaridade dos sentimentos dos tweets (positivo, neutro, negativo)

através dos Métodos de Análise de Sentimentos e Classificadores mais utilizados no mercado, como Naive Bayes (NB), e Máquinas de Vetores de Suporte (SVM), apresentando os principais benefícios de cada um, além das vantagens e desvantagens de um com relação ao outro em determinadas situações. Pretende-se também aplicar os métodos de análise antes e após o pré-processamento, com o intuito de mostrar em números a importância desta tarefa em meio ao processo como um todo.

## 1.3 Estrutura

Este trabalho está dividido em quatro principais capítulos de desenvolvimento. O capítulo 2 contém conceitos gerais, onde, apresenta-se o Twitter, identificando quais suas principais funcionalidades, características e peculiaridades, além do conceito geral de Big Data, detalhando o Ecossistema Hadoop.

O capítulo 3 apresenta os procedimentos necessários para efetuar a tarefa de análise de sentimentos, dando uma descrição do que se trata tal análise e descrevendo a coleta e armazenamento de dados. Em seguida, apresentam-se, métodos de pré-processamento de dados e, finalmente, são apresentadas as técnicas de análise de sentimentos propriamente ditas, divididas em supervisionadas e não supervisionadas.

O capítulo 4 apresenta a importância da tarefa de pré-processamento por meio de experimentos comparativos, além de se descreverem comparações entre as técnicas e métodos de análise apresentados.

Finalmente, o capítulo 5 contém as principais conclusões decorrentes deste trabalho.

## 1.4 Metodologia

Este trabalho apresenta discussões sobre processos, métodos e técnicas, apresentando uma análise e tendências sobre o tema escolhido, contendo métodos e técnicas de coleta de dados do Twitter, apresentando, no apêndice, os códigos desenvolvidos nesta monografia, para coletar tais dados e armazenar no Hadoop. Em seguida são apresentados métodos de pré-processamento encontrados em artigos científicos disponíveis no mercado, expondo apenas o conceito teórico dos mesmos, isto é, não foram colocados em prática neste trabalho. Da mesma forma, diversos

métodos de Análise de Sentimentos são levantados no decorrer deste, com intuito de apresentar os conceitos teóricos, ilustrando, através de experimentos disponíveis no mercado, os efeitos práticos dos mesmos.

# 2 CONTEXTUALIZAÇÃO

## 2.1 Twitter

O Twitter é uma RSO que permite que os usuários cadastrados postem mensagens curtas, de no máximo 140 caracteres, em tempo real. Os usuários que não possuem uma conta cadastrada podem apenas ler os chamados “tweets”. Esta RSO possuía em 2014, segundo França (2014), mais de 600 milhões de usuários, recebendo mais de 500 milhões de mensagens por dia e tinha uma média de 271 milhões de usuários ativos por mês. A rede social anunciou, em junho de 2017, o total de 328 milhões de usuários ativos. No Twitter, as redes podem ser formadas observando quem segue quem, quem mencionou quem ou quem fez um retweet (republicou a mensagem) de quem.

Em 2006, quando nascia o Twitter, ele era considerado como apenas mais uma rede social. Entretanto, com ele surgia uma nova forma de comunicação na Internet, pela qual as pessoas podiam divulgar qualquer tipo de informação, em tempo real, para todos aqueles ligados à sua rede. Conforme destacado por Nascimento; Osiek e Xexeo (2015), o limite de 140 caracteres soava como uma restrição de plataforma, mas esta característica determinou fortemente a forma como esta rede social é usada e o modo como os usuários se expressam nesta rede. Segundo Zhao e Rosson apud Nascimento; Osiek e Xexeo (2015), o fato de ser em tempo real, rápido e fácil publicar, fez com que as pessoas interagissem mais com a ferramenta. Tais características possibilitaram que esta rede tenha como base a troca de informações, de modo que o dado transmitido é a opinião dos usuários.

Ainda segundo Nascimento; Osiek e Xexeo (2015), essa RSO não se manteve apenas como mais uma rede social por dois grandes motivos:

1) O alto nível de popularidade atingido: o que demonstra a força dessa rede como fonte geradora de conteúdo subjetivo na Internet;
2) A forma como os usuários passaram a utilizá-la: não apenas divulgando informações sobre si, que era o seu intuito inicial, mas também compartilhando opiniões e informações sobre fatos e eventos em geral.

Tais fatos fazem do Twitter um feed de notícias baseado em pessoas, o que atrai, de modo geral, dois perfis de usuários: quem busca opiniões/informações de terceiros e aqueles que querem divulgar as suas. Deste modo, pode-se dizer que o

Twitter é uma importante fonte de opiniões e sentimentos que podem ser analisados e usados em diversas áreas, assim como será visto no decorrer deste trabalho.

Conforme Li e Li apud Nascimento; Osiek e Xexeo (2015) está destacada a importância de avaliar o sentimento dos tweets, pois identificaram que 20% deles estão relacionados a marcas e expressam opinião sobre a empresa ou produto. Além disso, Stelzner (2012) apontou que, em 2012, 83% dos profissionais de marketing indicavam que as mídias sociais são importantes para seus negócios.

Um fato interessante apontado por Nascimento; Osiek e Xexeo (2015) é a importância do compartilhamento de opiniões em tempo real,

> *"O fato de o Twitter permitir o compartilhamento em tempo real faz com que essa ferramenta permita captar o sentimento do usuário no momento em que ele soube da notícia em questão, o que o motiva a expressar sua emoção antes que outros fatores o influenciem e diminuam a intensidade do sentimento gerado."*

Por conta dos fatores apontados, o Twitter ganhou altas taxas de popularidade, disponibilizando informações valiosas para governos e negócios em todo o mundo. Tais informações potenciais disponíveis no Twitter são dados textuais que podem ser facilmente acessados e limpos. Os dados textuais podem ser usados para análises sentimentais do público em relação a um produto específico ou pessoa particular.

Segundo, França (2014), esta RSO armazena, além dos tweets, a identificação da mensagem, a data da postagem, localização geográfica do usuário que faz uma postagem (quando o usuário habilita), suas interações na rede, seus seguidores e pessoas que segue. Diante de todos esses dados disponibilizados, algo importante a ser observado é a estrutura dos mesmos. O trabalho de França (2014) destaca que, tratando-se de estrutura, as informações podem ser armazenadas de forma semiestruturada, isto é, com uma parte em estruturas/formatos e tipos pré-definidos, enquanto outra parte não. Já do ponto de vista de relacionamento de dados de diferentes fontes, as dificuldades são maiores, por exemplo, na identificação, relacionamento e análise de conteúdo de perfis dos usuários.

## 2.2 Big Data

Supraja; Mujamdar e Analaki (2015) definem Big Data como conjuntos de dados cujo tamanho está além da capacidade de bancos de dados típicos capturar, armazenar, gerenciar e analisar.

Esses conjuntos de dados podem ser gerados através de dispositivos móveis, redes de sensores, sistemas empresariais e o uso da internet, o que envolve as redes sociais, como o Twitter. O termo Big Data costuma ser associado a 3Vs: i) Volume, que diz respeito a um grande conjunto de dados; ii) Velocidade, referindo-se a necessidade de processamento rápido dos dados; e iii) Variedade, que está relacionada às distintas e diversas fontes de dados. Os grandes dados gerados podem ser estruturados, não estruturados ou semiestruturados.

França (2014) afirma que é importante observar também o Valor dos dados e as fontes de geração dos mesmos (Veracidade), isto é, dois outros “Vs” relacionados ao Big Data.

Com a modernização das tecnologias, tais como dispositivos, máquinas, sensores incorporados em veículos e o aumento do uso da internet, especialmente das redes sociais, como o Twitter, a quantidade de dados não estruturados gerados aumentou de forma exponencial. Essa grande quantidade de dados gerados a cada instante hoje é chamada de Big Data. O avanço da tecnologia tornou mais fácil a coleta e processamento de Big Data, que ainda crescerá muito num futuro próximo, principalmente por conta das redes sociais.

Diversas tecnologias utilizadas em Big Data, tais como bancos de dados NoSQL, Hadoop, Hive, Pig, têm como objetivo coletar, processar e armazenar dados importantes de maneira mais barata e eficiente. Big Data não diz respeito apenas ao armazenamento de grandes quantidades de dados, mas também a análise e a previsão de padrões ou tendências.

O Twitter, com milhões de usuários espalhados por todo o mundo e recebendo mensagens em alta frequência é um exemplo a ser observado. A quantidade de tweets enviados por usuários distintos gera um alto volume de dados e cada dado precisa ser armazenado, processado, relacionado a outras informações (seguidores, responsável pelo post), disponibilizado e publicado para outros usuários dessa mídia.

Costa (2012) discute a respeito do ciclo de vida dos dados, comparando com o ciclo de vida biológico, observando que um ciclo de vida pode ser classificado em quatro estágios:

- Nascimento (geração);
- Crescimento (agregação), ou seja, dados com semântica semelhante, ou mesmo dados que sejam de alguma forma correlacionados, são adicionados

aos dados originais, enriquecendo o valor dos dados e ampliando a sua importância;

- Reprodução (análise), onde a combinação dos dados obtidos resulta em novos dados com significado melhor e mais acurado;
- Morte (apagamento).

Outra comparação entre os ciclos de vida feita por Costa (2012) está relacionada ao estágio de crescimento, que diz respeito à movimentação dos dados de um local para outro, visando melhores ferramentas e condições de análise. No caso dos dados, diferente do ciclo de vida biológico, eles podem ser, além de migrados, replicados.

França (2014) faz um observação importante a respeito da fase de apagamento dos dados, afirmando não ser uma tarefa tão simples, pois é complexo definir quando certos dados não possuem mais valor para serem analisados.

> *Esse valor pode ser finalizado em um contexto, mas sob outros pontos de vista os dados podem possuir valor em novas análises.[..]*
> *Finalmente, não é possível definir valores fixos (prazos ou períodos exatos) de validade dos dados. Cabe aquele que gerencia o dado tomar a decisão de descartá-lo ou não. É um consenso que sempre que possível os dados devem ser mantidos (ou seja, a sua remoção deve ser evitada. (FRANÇA, 2014, p.26)*

Sabe-se que as tecnologias precisam atender os requisitos de armazenamento e processamento de Big Data (alta velocidade, capacidade de lidar com dados não relacionais, etc). Tratando-se do armazenamento, os Sistemas de Gerenciamento de Banco de Dados (SGBDs) tradicionais não conseguem lidar com volumes tão altos de dados. Tais volumes de dados requerem SGBDs capazes de processar dados estruturados e não estruturados, distribuindo-os.

Visto que é preciso explorar plenamente as informações disponíveis nas redes, tem-se a necessidade de identificar as diversas tecnologias que estão sendo desenvolvidas e adaptadas para manipular, analisar e visualizar Big Data. Deste modo, aborda-se neste trabalho, o Hadoop, que possibilita que aplicações escaláveis sejam desenvolvidas capacitando meios de processar os dados de forma distribuída e paralela.

A requisição de análise e gerenciamento de Big Data com alto desempenho vem aumentando muito por parte das empresas, considerando-se que está ficando cada vez mais comum a tarefa de análise de sentimentos dos dados de redes sociais

e, desse modo, diferentes soluções têm surgido. O paradigma MapReduce implementado pelo Hadoop permite o processamento distribuído de grandes volumes de dados, sendo muito vantajoso na construção de aplicações paralelas, com o intuito de processar dados estruturados e não estruturados em grandes escalas.

O MapReduce pode ajudar na otimização de diversos algoritmos de mineração de dados (utilizados na descoberta automática de modelos e padrões que usam técnicas como classificação, associação, regressão e análise de agrupamento), permitindo que sejam paralelizados, conforme descrito na seção 2.2.

### 2.2.1 Hadoop

Segundo Apache Hadoop (2017), o Hadoop pode ser definido como uma estrutura que permite o processamento distribuído de grandes conjuntos de dados em clusters de computadores usando modelos de programação simples. Em vez de confiar no hardware para oferecer alta disponibilidade, a própria biblioteca é projetada para detectar e lidar com falhas, oferecendo assim um serviço altamente disponível num conjunto de computadores, cada um dos quais pode ser propenso a falhas. Em outras palavras, o Hadoop é uma plataforma de código aberto eficiente e escalável que processa grandes quantidades de dados, de forma distribuída, isto é, pode haver diversos componentes distribuídos trabalhando conjuntamente para realizar uma única tarefa.

Geralmente, os grandes arquivos são distribuídos em vários clusters conhecidos como clusters HDFS (Hadoop Distributed File System). O HDFS é capaz de armazenar grande quantidade de dados sem perdas ou interrupções, além de gerenciar os clusters, dividindo os arquivos recebidos em pedaços chamados de pequenos blocos.

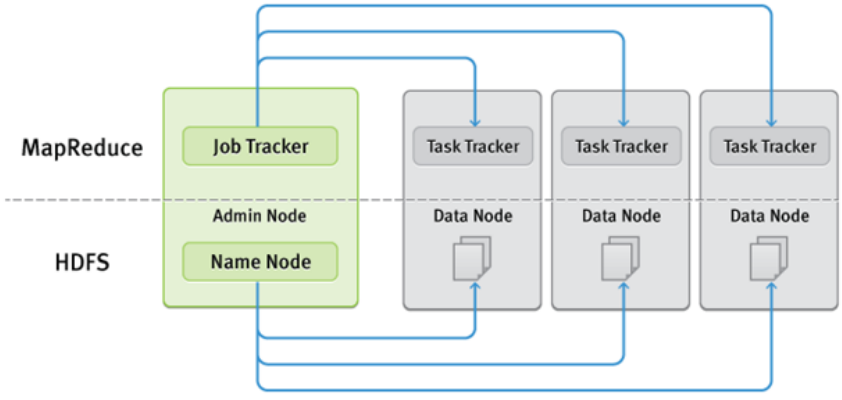
Segundo Apache Hadoop (2017), o HDFS consiste em um Namenode e Datanodes que gerenciam o armazenamento de blocos de dados. Namenode gerencia todos os datanodes e mantém seus metadados. O HDFS armazena cada arquivo como uma sequência de blocos. Esses blocos são replicados em vários nós e racks no HDFS, implementando a tolerância a falhas. Racks são compostos por uma coleção de datanodes que pertencem à mesma rede. Se um dos datanodes falhar, a réplica deste datanode, que está presente em outro nó, se move para um outro datanode.

O Hadoop é orientado para armazenar e processar vastos volumes de dados e implementa uma estrutura chamada MapReduce. As consultas são divididas entre diferentes nós, a serem realizadas em paralelo e isso é conhecido como o estágio de *Mapeamento*. Em seguida, os resultados são combinados no estágio de *Redução* para retornar uma saída. Isso fornece uma execução de consultas e uma provisão de resultados mais rápidos. MapReduce lida com processamento paralelo em nós, com a ajuda de um JobTracker (no master) e TaskTrackers (nós escravos).

Apache Hadoop (2013) descreve o processo de MapReduce como uma estrutura de software para escrever aplicativos que processam grandes quantidades de dados (conjuntos de dados de vários terabytes) em paralelo em grandes clusters (milhares de nós) de maneira confiável e tolerante a falhas. O MapReduce geralmente divide o conjunto de dados de entrada em blocos independentes que são processados pelas tarefas do Map de forma completamente paralela. A estrutura classifica as saídas dos Maps, que são então inseridas nas tarefas de Reduce. Normalmente, tanto a entrada como a saída são armazenadas em um sistema de arquivos. A estrutura MapReduce possui um único mestre JobTracker e um escravo TaskTracker por nó. O mestre é responsável pelo agendamento das tarefas dos escravos, de modo que estes executam as tarefas conforme o mestre.

A Figura 1 mostra o funcionamento do HDFS junto ao processo de MapReduce:

**Figura 1 - Funcionamento HDFS e MapReduce**




**Fonte: França (2014)**

Segundo França (2014), o Hadoop apresenta três grandes benefícios para a análise de redes sociais:

i) Capacidade de armazenar grandes volumes de dados utilizando máquinas de commodity (dispositivo ou componente relativamente barato e disponível);

ii) Armazenamento de dados com formatos variados;

iii) Modelo de processamento paralelo de alto nível.

Existem diversas ferramentas e bibliotecas que auxiliam em tarefas administrativas para o cluster, no processamento e análise de dados e no próprio armazenamento de dados do Hadoop. Tais ferramentas e bibliotecas são chamadas de Ecossistema Hadoop. França (2014) listou algumas destas ferramentas e bibliotecas, conforme descrito a seguir:

- Tableau: plataforma de visualização e análise de dados. Possui versões gratuitas para estudantes que podem ser encontradas em http://www.tableausoftware.com/pt-br;
- R: ambiente para análises estatísticas, disponível em http://www.r-project.org/;
- Java: Java é a linguagem oficial para criar programas em um cluster Hadoop;
- Pig: é uma plataforma para análise de dados que, segundo Apache Hadoop (2017) consiste de uma linguagem de alto nível para expressar uma análise de dados e a infraestrutura para executar essa linguagem;
- Hive: uma infraestrutura que fornece um mecanismo para projetar, estruturar e consultar os dados usando uma linguagem baseada em SQL, chamado HiveQL, conforme Apache Hadoop (2017);
- Flume: conforme Apache Flume (2017), este é um serviço distribuído, confiável e disponível para coletar, agregar e mover de forma eficiente streams de dados;
- Sqoop: é uma ferramenta projetada para transferência eficiente de dados em massa entre o Apache Hadoop e bancos de dados estruturados, conforme Apache Sqoop (2018);
- Zoo Keeper: serviço de coordenação de alto desempenho para aplicações distribuídas, segundo Apache Hadoop (2017).

# 3 ETAPAS PARA A REALIZAÇÃO DA ANÁLISE DE SENTIMENTOS

## 3.1 Descrição

Segundo Jianqiang e Xiaolin (2017), a Análise de Sentimento é o trabalho de tratar computacionalmente a opinião ou sentimento público e também a subjetividade em um texto especialmente obtido das mídias sociais. Pode ser usada para conhecer a opinião dos usuários em relação a um tópico específico. Em alguns casos pode ser usada para julgar o sucesso ou o fracasso de um produto, marca, político ou mesmo celebridade.

Tais identificações de sentimentos são feitas de forma automática, atribuindo uma polaridade, por exemplo, a um tweet, isto é, determinando se o mesmo é positivo, negativo ou neutro, através de técnicas, algoritmos e ferramentas que serão abordados na seção 3.5.

A Análise de Sentimentos vem sendo usada cada vez mais pelas grandes empresas, coletando dados principalmente do Twitter, que, como dito, contém informações valiosas a respeito da opinião pública, podendo ser aproveitado da forma correta.

Segundo Agarwal et. al. (2011), a análise do sentimento foi tratada como uma tarefa de Processamento de Linguagem Natural (PLN) em muitos níveis de granularidade. Começando como uma tarefa de classificação no nível do documento, em seguido passou a ser tratada também no nível da sentença e mais recentemente no nível de frase, como no caso de análise de tweets.

## 3.2 Coleta dos Dados

O primeiro passo a ser dado no trabalho de análise de sentimentos consiste da coleta dos dados, isto é, a extração das informações, neste caso, do Twitter. É sabido que a maior parte dos dados da web não possui uma estrutura definida, ou, quando possui, ela atribui poucas informações a respeito dos dados presentes na página.

Chen (2001) estimou que 80% do conteúdo online do mundo em 2001 era baseado em texto, além disso, sabe-se que dados não estruturados contemplam não apenas textos, mas também imagens, vídeos e músicas, o que reafirma que dados não estruturados compõem a web.

A análise de sentimentos provoca uma necessidade de estruturação desses dados, mas este trabalho não é tão simples e deve ser vastamente estressado dado a grande variedade dos dados existentes na web. Segundo França (2014), alguns padrões foram criados com o intuito de solucionar este problema. Os padrões XML e JSON são os mais utilizados, mas existem diversos outros e nada impede de que novos sejam criados ou aplicados.

França (2014) apresentou duas grandes formas de coletar dados das redes sociais:

1) Coleta através de termos determinados, trazendo dados do passado, o que pode ter restrições, dado que geralmente existe um período de tempo viável para a coleta dos dados.

2) Coleta através de um conceito de streaming, no qual a aplicação criada funciona como um “ouvinte” da rede e captura os dados à medida que estes surgem.

França (2014) aponta que o Twitter disponibiliza duas APIs (Application Programming Interface) para capturar dados: Streaming API e REST API. A diferença entre elas é que Streamings APIs suportam conexão de longa duração e fornecem dados em tempo quase real, enquanto REST APIs suportam conexões de curta duração e são limitadas (pode-se baixar uma determinada quantidade de dados).

Para utilizá-las é necessário, primeiramente, que o usuário tenha uma conta no Twitter, com isso é possível acessar a página https://apps.twitter.com, autenticar-se com sua conta e criar uma nova aplicação. Após este cadastro será possível coletar o consumerKey, consumerSecret, accessToken e accessTokenSecret. Cada aplicação tem este grupo de tokens de acesso em nome de um usuário. Esses tokens são únicos para cada aplicação e cada usuário. Os dois primeiros estão relacionados com a aplicação enquanto os outros dois estão relacionados com o usuário que concorda em autorizar a aplicação a executar pedidos em seu nome. Além disso, o usuário tem a capacidade de revogar os tokens gerados a qualquer momento, o que resultará na desativação dos mesmos, restringindo a aplicação de executar pedidos de API em nome do usuário. Essas chaves serão utilizadas na autenticação da aplicação de captura de dados.

O Twitter trabalha com o padrão de arquivo JSON, de forma que todos os dados são recebidos neste formato. França (2014) implementou uma solução utilizando

Streaming API conforme código contido no Anexo I. Neste caso, o termo que a aplicação está “ouvindo”, trata-se do parâmetro “bom dia” passado no código. Os programas usados neste exemplo foram codificados em Python. França (2014) utilizou as bibliotecas “Auth1Session” e “JSON”. A Auth1Session estabelece a conexão com o Twitter e a biblioteca JSON transforma o texto recebido em um objeto Python com estrutura no formato JSON, de modo que é possível manipular o arquivo.

França (2014) também implementou uma solução utilizando a REST API do Twitter, conforme mostra o código disponibilizado no Anexo II. Neste caso a consulta foi feita pelo termo “israel” através da uma variável. A API retorna os 100 tweets mais recentes que contêm esse termo. Este limite de 100 foi passado na URL de requisição através de um parâmetro. Todos os termos que forem passados são retornados pela API. Em ouras palavras, se essa variável possui o termo “brasil futebol”, então serão retornados os tweets que possuem as duas palavras, como por exemplo “O Brasil é o país do futebol”.

França (2014) citou ainda uma restrição referente à REST API: o Twitter não permite que a API busque por tweets que tenham sido postados há mais de sete dias e ainda bloqueia a aplicação caso ultrapasse o número de requisições permitidas, sendo necessário um intervalo de 15 minutos para que a aplicação seja desbloqueada.

Nos últimos anos, as principais plataformas de Redes Sociais Online (RSO) usavam políticas baseadas em IP, restringindo uma máquina para executar certo número de solicitações. A solução direta para enfrentar este desafio era um procedimento de coleta de dados distribuído.

No entanto, as plataformas RSO, como o Twitter, alteraram estas políticas baseadas em IP para políticas baseadas em aplicação, de modo a restringir uma aplicação de realizar um grande número de solicitações. Esta atualização tornou o processo de coleta de dados em grande escala mais complicado.

O trabalho de Efstathiades et. al. (2016) descreveu uma solução eficiente para superar os desafios citados, referentes às limitações impostas pelas políticas de API do Twitter. Foi apresentado nele uma estrutura de coleta de dados Crowd Crawling, que permite que pesquisadores efetuem campanhas de coleta de dados em larga escala com a participação de usuários da rede. A solução proposta é capaz de

recolher dados históricos de forma assíncrona, assim como recuperar o fluxo em tempo real.

### 3.2.1 Crowd Crawling: Construindo o Repositório de Tokens

Conforme mencionado, um desafio introduzido nos procedimentos de coleta de dados foi a atualização das limitações baseadas em IP (Internet Protocol) para aqueles baseados em aplicações. Em limitações baseadas em IP, a API monitora o endereço de IP da máquina e aplica as limitações. Assim, uma campanha de coleta de dados distribuídos em diferentes máquinas melhora radicalmente o procedimento.

No entanto, após as atualizações das políticas baseadas em IP para políticas baseadas em aplicação, isso deixou de ser possível, mesmo quando uma aplicação (que utiliza o mesmo conjunto de tokens) é distribuída em várias máquinas com diferentes endereços IP, as limitações que se aplicam são as mesmas que se estivessem sendo executadas em uma única máquina. Desse modo, Efstathiades et. al. (2016) encontrou uma solução: pedir aos usuários da rede que contribuam para o procedimento de coleta de dados, autorizando as aplicações a acessar a API.

Segundo Efstathiades et. al. (2016) este é o procedimento sugerido pela plataforma RSO. Primeiro, registraram uma aplicação API no Twitter, então, desenvolveram um serviço que pede aos usuários da rede de amigos (seguintes e seguidores) para autorizar a execução de pedidos de recuperação de dados públicos. Tendo a aprovação do usuário, o serviço coleta os tokens gerados e os armazena em um Repositório de Tokens. Este repositório contém uma série de tokens que foram gerados pelos usuários RSO. Este procedimento, que foi chamado de crowd crawling, aumenta o número de tokens/recursos que podem ser usados durante o processo de recuperação no sistema proposto e ocorre antes do início da campanha de coleta de dados.

Ter vários tokens permite ativar um conjunto diferente deles para evitar atingir o limite de solicitação. Quando o limite é atingido, o grupo de tokens torna-se inválido por um certo período de tempo “t”. Em seguida, passa-se para o próximo grupo de tokens até este atingir o limite também. Este procedimento é seguido até que o tempo “t” tenha passado, de modo que o grupo de tokens inicial se tornará ativo novamente. Assim, com um número “n” de tokens, é possível habilitar a operação contínua da campanha de coleta de dados.

O texto Efstathiades et. al. (2016) entra no detalhe dessa solução, explicando cada componente e seu papel, mostrando como deve funcionar para coleta de dados específicos e coleta de transmissão em tempo real.

## 3.3 Armazenamento

Como apresentado, a ascensão exponencial do uso da internet reflete-se no crescimento referente à geração de dados nas RSO. As empresas responsáveis pelas RSO possuem Data Centers capazes de manter essa gigantesca quantidade de dados, graças ao retorno financeiro que tais redes proporcionam. Diversas empresas fornecem as tecnologias necessárias para tratar grandes volumes de dados, além de algumas RSOs que já possuem condições de criar suas próprias tecnologias para armazenar e gerenciar esse grande volume de dados.

Quando há pretensão de coletar, armazenar, gerenciar e analisar tais dados esbarra-se na necessidade de excluir partes dos dados ou mesmo armazenar em tecnologias de menor custo. Entretanto, tais tecnologias mais baratas tornam o acesso aos dados mais difícil. Segundo França (2014),

> Quando se trata da análise, outros fatores se somam ao volume, como o tipo de análise que é realizada: grafos com milhões de nós e de centenas de milhões ou bilhões de áreas, processamento de linguagem natural de textos diferentes que exigem grande quantidade de processamento, entre outros exemplos. Nesses casos a dificuldade é saber qual (ou quais) plataforma(s) de hardware se deve utilizar para lidar com grandes massas de dados, os quais superam a capacidade de tratamento possibilitada pelos sistemas tradicionais. (FRANÇA, 2014, p. 29)

Diante destas dificuldades uma solução que tem sido amplamente adotada pelas empresas ou pesquisadores que desejam coletar e analisar esses dados é distribuir e paralelizar o processamento e armazenamento dos dados que desejam manipular em cluster usando soluções como o Hadoop. No caso de uma solução como esta, uma infraestrutura já existente e que passa a não ser suficiente diante do novo cenário pode ser aproveitada, redirecionando os recursos subutilizados de hardware para o processamento e armazenamento dos dados.

### 3.3.1 Exemplo prático de captura e armazenamento

Detalha-se nesta seção como pode ocorrer a captura completa de tweets utilizando o Streaming API através do Flume. Neste caso, usa-se o Flume para o lançamento de dados do Twitter no HDFS do Hadoop.

Para uso do Flume na captura dos dados do Twitter, dentro da Cloudera (empresa que fornece a distribuição Apache Hadoop, além de suporte, treinamento e serviços profissionais) o primeiro passo é criar o usuário "Flume" e seu “Home” no HDFS, onde os dados serão armazenados. Isso pode ser feito através do User Admin do aplicativo.

Feito isso, é preciso baixar o .jar do Flume para JSON, que pode ser encontrado em https://github.com/mapr/mapr-demos/raw/master/drill-twitter-MSTR/flume/target/original-flume-sources-twitter-json-0.1.jar e, em seguida, executar os scripts disponíveis no Apêndice I, que mostram os passos:

1) Iniciar os serviços Cloudera;
2) Ir ao diretório do Flume;
3) Criar o arquivo “/home/training/twitter.conf”;
4) Subir o serviço do Flume;
5) Criar a tabela;
6) Selecionar os campos.

O arquivo mencionado no terceiro passo configura o HDFS como o repositório, define o caminho para armazenar tweets no mesmo, configura o termo que deverá ser “escutado” e parametriza os tokens do Twitter.

Executando o arquivo de configuração, os tweets começarão a baixar no HDFS no caminho especificado, para isso, basta subir o serviço do Flume, conforme descrito no quarto passo do Apêndice I. Após alguns minutos, os tweets que contêm o termo definido no arquivo de configuração devem aparecer no HDFS.

Os dados baixados no HDFS estarão no formato JSON e será preciso convertê-los em um formato legível. Os passos 5 e 6 do Apêndice I mostram os scripts de criação de tabela e de seleção de campos, fazendo um “catado” de alguns dos atributos disponibilizados pelo Twitter, tornando os dados legíveis de forma estruturada. Para tal, pode-se utilizar o HUE (Hadoop User Experience) que é, segundo Hue (2013) uma interface de código aberto baseada em Web que torna o Apache Hadoop mais fácil de usar, permitindo uma interação com os serviços Hadoop sem ter que ir a uma interface de linha de comando. O HUE possui diferentes aplicações, dentre elas um editor para o Apache Hive. Aberto o HUE, basta criar a tabela apontando para o diretório dos tweets e em seguida, selecionar os campos desejados utilizando a função getjson.

Feito isto, passa a ser possível consultar os dados de forma estruturada, o que permite que os mesmos sejam analisados. Entretanto, outra tarefa indispensável nesse contexto se trata do pré-processamento destes dados, o que será visto em detalhes na seção 3.4.

### 3.4 Pré-processamento

Nesta fase, os tweets estão disponíveis como dados de texto e cada linha contém um tweet. Pode-se achar que esta etapa do processo não possui tanta importância, mas mostra-se neste trabalho o quanto os resultados das análises podem variar de acordo com a escolha do método de pré-processamento. Os Tweets geralmente são compostos por frases incompletas, poluídas e mal estruturadas, contendo expressões irregulares e palavras que não existem no dicionário, por isso é preciso tomar cuidado ao selecionar os métodos que melhor se adequem e mais agreguem ao processo como um todo.

Segundo Jianqiang e Xiaolin (2017), a maioria das abordagens existentes para identificar a polaridade do sentimento dos tweets aplica o pré-processamento no texto (por exemplo, remover URLs, amplificação de siglas, substituição de menções negativas, remoção de palavras de parada, etc) para reduzir a quantidade de ruído dos tweets. A hipótese é que o pré-processamento de dados reduz o ruído, ajuda a melhorar o desempenho do classificador e acelerar o processo de classificação.

Ainda conforme Jianqiang e Xiaolin (2017), há uma falta de análise adequada e profunda do impacto do pré-processamento de texto na classificação de sentimento do Twitter. Para preencher essa lacuna, este mesmo autor detalhou os principais métodos de pré-processamento utilizados na limpeza de tweets. Expõe-se, a seguir, cada uma delas para que na seção de análise/comparação seja possível discutir quando cada método se adequa mais:

- Substituição de menções negativas (este método faz sentido quando se tratam de tweets escritos em Inglês). Os Tweets consistem em várias noções de negação. Em geral, a negação desempenha um papel importante na determinação do sentimento do tweet. Neste caso, o processo de negação transformando "won't", "can't", e "n't" em "will not", "cannot", and "not", respectivamente.

- Remoção de links de URL (Uniform Resource Locator). Aqui, os URLs curtos do Twitter são expandidos e são tokenizados. Em seguida, o URL é substituído pelo token no tweet para refinar o conteúdo do mesmo.
- Rompimento de palavras que contêm letras repetidas à sua forma original. As palavras com letras repetidas, por exemplo, "muuuuuito", são comuns em tweets, e as pessoas tendem a usar dessa maneira para expressar seus sentimentos. Aqui, uma sequência de mais de três caracteres semelhantes é substituída por três caracteres. Por exemplo, "muuuuuito" é substituído por "muuuito". O uso de três caracteres destaca quando há uma entoação mais intensa nessa palavra.
- Remoção de números. Em geral, os números não servem de nada ao medir o sentimento e são removidos dos tweets para refinar o conteúdo do mesmo.
- Remoção de palavras de parada. Palavras de parada geralmente se referem às palavras mais comuns em um idioma, como "o", "de" e "para". A maioria dos pesquisadores considera que as palavras de parada desempenham um papel negativo na tarefa de classificação de sentimento e são removidas antes da seleção de recursos pelos pesquisadores. O método clássico de remover palavras de parada é o método baseado em listas pré-compiladas. Existem várias listas prontos disponíveis na rede.
- Ampliação de siglas ou abreviações para as palavras originais usando um dicionário de siglas. Gírias, abreviações e siglas são comuns nos tweets, por isso é necessário expandi-las. Para tal tarefa também existem dicionários prontos disponíveis na Internet.
- Stemming. Refere-se a identificar a raiz de uma palavra. Por exemplo, as palavras “working”, “work“ e “worked“ não possuem distinção de polaridade, assim é possível reduzi-las à sua raiz “work“.

## 3.5 Técnicas de Análise de Sentimentos

A análise do sentimento é um processo que pode ser aplicado em nível de documento, nível de parágrafo ou nível de sentença, com o intuito de descobrir se a polaridade é positiva, negativa ou neutra. Conforme já descrito, ela visa descobrir a opinião de um determinado público em relação a um assunto de interesse.

Nesta seção passa-se uma visão mais aprofundada a respeito das técnicas e estratégias que existem hoje no contexto de análise de sentimentos e suas principais aplicações no âmbito das redes sociais online.

Como este tema se popularizou nos últimos anos, diversos termos e conceitos estão sendo caracterizados diante de cada nova atividade associada a obtenção de sentimentos. Benevenuto et al. (2015) detalhou alguns destes termos:

**Polaridade**: é o grau que representa o quanto um texto é positivo ou negativo. Esta costuma ser uma saída para os métodos de análise de sentimentos. Alguns métodos tratam a polaridade como um resultado binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro). Por exemplo, a frase “O meu dia foi maravilhoso” é positiva, a frase “Eu odeio este país” é negativa e a frase “Hoje é 17 de fevereiro” é classificada como neutra.

**Força do sentimento**: Representa a intensidade de um sentimento ou da polaridade, sendo também uma forma de saída de alguns métodos. Normalmente é um valor flutuante entre (-1 e 1) ou até entre ($-\infty$ e $+\infty$).

**Sentimento/Emoção**: Indica um sentimento específico presente em uma mensagem (raiva, surpresa, felicidade, etc.). Alguns métodos apresentam abordagens capazes de identificar qual sentimento em específico uma sentença representa.

Além destes termos, Benevenuto et al. (2015) destaca que as técnicas de Análise de Sentimentos podem ser classificadas como Supervisionadas ou Não Supervisionadas.

### 3.5.1 Técnicas não supervisionadas

As Técnicas Não Supervisionadas (TNS) não necessitam de sentenças previamente rotuladas ou treinos para a criação de um modelo, sendo esta uma das suas principais vantagens, dado que não são exclusivas a um determinado contexto.

Conforme Benevenuto et al. (2015), dentre as técnicas não supervisionadas, pode-se destacar aquelas com abordagens léxicas, embasadas em um dicionário léxico de sentimento, como se fosse um dicionário de palavras, entretanto que possui no lugar do significado das palavras, um significado quantitativo ou qualitativo.

Um léxico de sentimento abrange listas de expressões e palavras usadas para expressar sentimentos e opiniões subjetivas das pessoas. Por exemplo, utilizando

um léxico de palavras positivas e negativas, basta analisar o documento para o qual o sentimento precisa encontrar, se o documento tiver mais palavras de léxico positivas, é positivo, caso contrário, é negativo.

As técnicas baseadas em léxico para a análise de sentimentos são a aprendizagem sem supervisão porque não requerem treinamento prévio para classificar os dados. Os passos básicos das técnicas baseadas em léxico foram descrito por Vohra e Teraiya (2013) conforme apresentado a seguir:

1. Pré-processe cada texto (conforme descrito na seção 3.4);

2. Inicialize a pontuação total de sentimentos do texto: s = 0.

3. Tokenize o texto. Para cada token, verifique se ele está presente em um dicionário de sentimentos.

(a) Se o token estiver presente no dicionário:

i. Se o token for positivo, então s = s + w (onde w é o valor que representa a Força do Sentimento).

ii. Se o token for negativo, então s = s - w.

4. Observe a pontuação total do sentimento do texto s,

(A) Se s > 0, então o texto é classificado como positivo.

(b) Se s < 0, o texto é classificado como negativo.

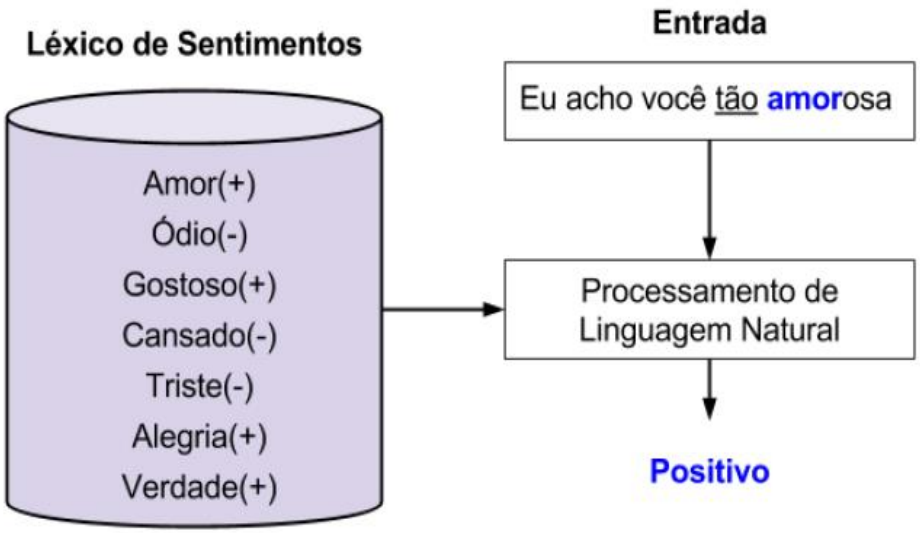
Segundo Vohra e Teraiya (2013), existem três métodos para construir um léxico de sentimento: construção manual, métodos baseados em corpus e métodos baseados em dicionário. A construção manual do léxico do sentimento é uma tarefa difícil e demorada, dado que ele deve conter uma grande quantidade de termos, que devem estar criteriosamente divididos.

Em técnicas baseadas em dicionário, a ideia é primeiro coletar um pequeno conjunto de palavras de opinião manualmente com orientações conhecidas e, em seguida, aumentar esse conjunto pesquisando no dicionário WordNet seus sinônimos e antônimos. As palavras recém-descobertas são adicionadas à lista.

As técnicas baseadas em Corpus dependem de padrões sintáticos em corpos grandes. Estas técnicas podem produzir palavras de opinião com uma precisão relativamente alta. A maioria desses métodos precisa de uma grande quantidade de

dados de treinamento rotulados. Esta abordagem tem uma grande vantagem que a abordagem baseada em dicionário não possui. Pode ajudar a encontrar palavras de opinião específicas do domínio e suas orientações. Em outras palavras, estas abordagens não contam com a preparação de modelos de Machine Learning e, em geral, são oriundas de tratamentos léxicos de sentimentos que abrangem, de forma macro, calcular a polaridade de um texto ou sentença, em cima da orientação semântica dos termos presentes no mesmo.

**Figura 2 - Léxico de Sentimentos**




Fonte: Benevenuto et al. (2015)

A Figura 2 demonstra, de modo geral, como funciona o método de análise de sentimentos léxico. A classificação se inicia quando uma sentença de entrada é disponibilizada, onde ocorre o processamento de linguagem natural, em seguida é feita a busca dos termos que constituem esta mensagem no léxico. Finalizado este procedimento, é possível identificar a polaridade ou sentimento contido no texto de entrada.

Existem diversos dicionários prontos na literatura, sendo alguns deles mais conhecidos e completos, como LIWC (Linguistic Inquiry and Word Count), General Inquirer e Opinio Lexicon que são classificados como Binários (Positivo/Negativo), além de ANEW (Affective Norms for English Words ), SentiWordNet e SenticNet, que trabalham com Intensidade do Sentimento (entre -1 e 1).

A construção de um dicionário léxico pode ser muito complexa, dado que existe um alto volume de dados gerados na diariamente na web. Esses dados podem

variar de uma linguagem muito formal até uma totalmente coloquial e um dicionário completo deve contemplar todos os cenários possíveis. Neste caso, como se tratam de dados coletados do Twitter, espera-se que as expressões estejam numa linguagem pouco formal, contendo sarcasmo, ironia, gírias e palavras de baixo calão, onde diariamente podem surgir novos jargões, memes e hashtags, os quais devem ser corretamente compreendidos.

Para que a abordagem léxica seja funcional é necessário que ocorra a etapa de pré-processamento já descrita.

> Existe uma linha de pesquisa cujo estudo foca exatamente neste aspecto, chamada de Processamento de Linguagem Natural (PLN) e envolve basicamente o estudo e a compressão por computadores de como humanos naturalmente falam, escrevem e se comunicam. Obviamente os computadores não possuem a mesma capacidade de interpretação que os humanos e necessitam de algoritmos precisos e não ambíguos para serem capazes de realizar tal tarefa. (BENEVENUTO et al., 2015, p. 12)

Benevenuto et al. (2015) também exemplifica como os computadores dividem as partes de um texto em elementos gramaticais. Na frase: “The amazing Cloud delivers data to me ASAP”, o processamento computacional quebra as palavras em elementos gramaticais (“amazing” = adjetivo; “cloud” = substantivo; “delivers” = verbo), identificando que “cloud” se refere a “cloud computing” e “ASAP” a uma abreviação para “As Soon As Possible”. Esta tarefa de dividir o texto em elementos gramaticais é chamada de POS (Part-of-Speech) e existem algumas bibliotecas disponíveis capazes de fazê-la. Tais informações a respeito do texto são fundamentais para a análise de sentimento, dado que a mudança na característica gramatical de uma palavra pode alterar o que ela significa e a intensidade do sentimento envolvido na mesma.

Benevenuto et al. (2015) ressalta também que um dicionário léxico não é capaz de retornar a classificação das sentenças de maneira eficaz sozinho, dado que o simples somatório da pontuação de cada uma das palavras pode apresentar resultados superficiais. Por outro lado, muitos métodos se baseiam nas polaridades previamente definidas por dicionários léxicos juntamente com outras heurísticas e processamentos que possibilitam maior efetividade na identificação do sentimento das sentenças.

Conforme Benevenuto et al. (2015), tais heurísticas referem-se a detalhes da gramática que podem alterar a intensidade do sentimento e são mais do que somatórias de pontuações. Por exemplo, o número de exclamações ao final de uma

sentença, frases escritas em letras maiúsculas ou mesmo uma conjunção "mas" que pode mudar a polaridade da frase.

Os dicionários podem classificar as sentenças de duas formas: determinando se a sentença é positiva, negativa ou neutra, ou definindo a intensidade do sentimento (entre –x e x, sendo x o valor atribuído à Força do Sentimento). Dentre os principais dicionários que atuam do primeiro modo, detalha-se a seguir o LIWC e dentre aqueles do segundo modo, o SenticNet.

**a) LIWC**

Segundo Tausczik e Pennebaker (2010), o LIWC é uma ferramenta que possui um dicionário léxico com cerca de 4500 palavras classificadas em oitenta categorias. A palavra “amor”, por exemplo, está classificada na categoria “emoções positivas”, podendo, cada palavra, estar em mais de uma categoria. Tais categorias foram decididas e alimentadas através de buscas em dicionários, questionários e listas desenvolvidas por pesquisadores.

Tausczik e Pennebaker (2010) descrevem que LIWC tem duas características centrais: o componente de processamento e os dicionários. O recurso de processamento é o próprio programa, que abre uma série de arquivos de texto e depois passa palavra por palavra em todo o arquivo. Cada palavra em um determinado arquivo de texto é comparada com o arquivo de dicionário. Por exemplo, se LIWC estivesse analisando a frase “foi um verão quente e chuvoso” o programa primeiro olharia a palavra "foi" e verificaria se está no dicionário, identificando às categorias de verbos, verbos auxiliares e verbos do passado. Em seguida, a palavra "um" seria verificada e associada à suas categorias. Depois de passar por todas as palavras, LIWC calcula a porcentagem de cada categoria. Assim, pode-se descobrir que uma porcentagem X de todas as palavras de um determinado tweet, por exemplo, são palavras de emoção negativa e uma porcentagem Y de emoção positiva. Assim, LIWC lista todas as categorias e as taxas que cada categoria foi usada no texto fornecido.

**b) SenticNet**

Segundo Benevenuto et al. (2015), SenticNet é um dicionário semântico e afetivo que realiza a análise de sentimentos por meio dos conceitos das palavras. Ele foi desenvolvido através de sentic computing: um paradigma que explora as ciências da computação e sociais para melhor reconhecer, interpretar e processar opiniões e sentimentos na Web. Conforme, Cambria et al. (2010), ele possui cerca

de 14000 conceitos (como adoração e admiração), aos quais são associados valores de polaridade de sentimento em uma escala que varia de -1 até 1.

Benevenuto; Ribeiro e Araújo (2013) exemplificam o uso deste dicionário com a frase “Boring, it’s Monday morning”. Inicialmente SenticNet identifica os conceitos da frase: “boring” e “Monday morning”. Em seguida, calcula a polaridade para cada conceito, nesse caso -0,383 para “boring”, e +0,228 para “Monday morning”. O resultado final para sentimentos neste exemplo seria de -0,077, que é a média dos valores encontrados para cada conceito.

### 3.5.2 Técnicas Supervisionadas

As Técnicas Supervisionadas (TS) se baseiam em conceitos de Machine Learning, onde são necessários dois conjuntos de documentos: conjunto de treinamento e conjunto de teste. O conjunto de treinamento é usado por um classificador automático para aprender as características diferenciadoras dos documentos, e o conjunto de testes é usado para verificar o desempenho do classificador. Benevenuto et al. (2015) listou quatro etapas centrais para o processo de aprendizagem de máquina:

1. Obtenção de dados rotulados que serão utilizados para treino e para teste;
2. Definição de características que permitam a distinção entre os dados;
3. Treinamento de um modelo computacional com um algoritmo de aprendizagem;
4. Aplicação do modelo.

Técnicas de aprendizado de máquinas como Naive Bayes (NB), Entropia Máxima (ME), Logistic Regression (LR), Random Forest (RF) e Máquinas de Vetores de Suporte (SVM) alcançaram um grande sucesso na análise de sentimentos. Detalha-se a seguir, duas dessas técnicas.

***a)*** **Máquinas de Vetores de Suporte (MVS)**

Esta técnica vem ganhando muita atenção da comunidade de Aprendizado de Máquina. Segundo Gonçalves (2016) esta técnica funciona construindo um hiperplano ou um conjunto de hiperplanos capazes de fornecer modelos de classificação e regressão. Para as máquinas de vetores de suporte uma boa separação é constituída de uma maior margem, dada pela distância entre as observações (pontos de treino) e o hiperplano que separa a classe.

Lunardi; Viterbo e Bernardini (2016, p. 3) explicam que “em casos simples com dois grupos de dados, diz-se que os grupos são linearmente separáveis se existe um hiperplano que os separa”. Sabendo que, de um lado estão informações de uma classe e do outro lado informações de outra, diz-se que o hiperplano é o limitador de decisão. Segundo Lunardi; Viterbo e Bernardini (2016), o objetivo é localizar os pontos mais distantes da linha separadora. Conforme descrito em Harrington (2012), esses pontos que separam o hiperplano são conhecidos como vetores de suporte. Em contraponto, existem problemas que não são linearmente separáveis, assim, as SVMs não lineares identificam o conjunto de dados de treinamento do seu espaço inicial para um novo de maior dimensão, conhecido como espaço de características.

**b) Naive Bayes (NB)**

Naive Bayes é um classificador de polaridade de sentimentos probabilístico baseado em aprendizado de máquina. Segundo Jinturkar e Gotmare (2016) esse tipo de classificador é altamente escalável e pode lidar com vários parâmetros de forma eficaz. Conforme destaca Carvalho (2014) ele é bastante utilizado por ser rápido e simples de implementar e é visto como um dos mais eficientes em questões que dizem respeito a processamento e exatidão na classificação de novas amostras. Este método é baseado na aplicação do Teorema de Bayes, que tem a representação conforme Expressão (1), onde B representa um evento que ocorreu previamente e A um evento que depende de B. Para calcular a probabilidade de A ocorrer dado o evento B, o algoritmo conta o número de casos em que A e B ocorrem juntos e divide pelo número de casos em que B ocorre sozinho.

$$P(A|B) = \frac{P(B_1|A).P(B_2|A)...P(B_n|A).P(A)}{P(B_1).P(B_2)...P(B_n)} \qquad (1)$$

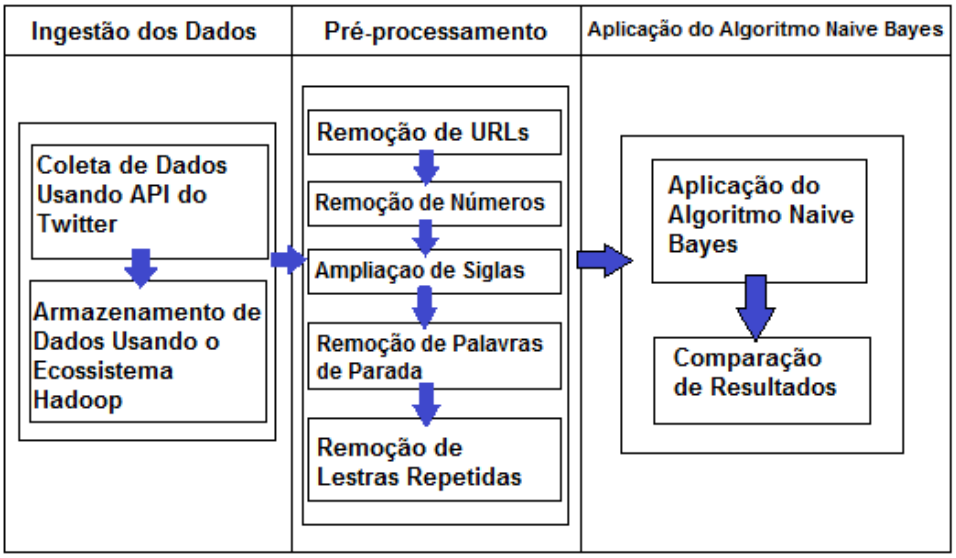
É importante descrever como o método Naive Bayes pode ser aplicado utilizando o ecossistema Hadoop apresentado. Parveen e Pandey (2016) apresentaram esta implementação de forma completa, onde o primeiro passo para tal

é a obtenção de um dicionário SentiWordNet treinado que esteja disponível on-line. Este dicionário consiste de coleções de palavras diferentes com o seu sinônimo e sua polaridade. É necessário inserir 2 arquivos o map: conjunto de dados do Twitter que contém os twittes dos usuário (conforme apresentado na seção 3.3) e o dicionário SentiWordNet que contém a polaridade das diferentes palavras. Em se tratando de aplicar este algoritmo utilizando o ecossistema Haddop, há duas fases na metodologia, Map e Reduce.

Conforme Parveen e Pandey (2016), na fase do mapeamento ocorrem duas tarefas principais: primeiro, cria-se um map hash para recuperar a polaridade de cada palavra, em seguida, deve-se processar a polaridade geral dos tweets aplicando o algoritmo Naive Bayes. O método map () na fase MapReduce lê o conteúdo do dicionário SentiwordNet a partir de um arquivo e transformando-o no mapa Hash para a recuperação de polaridade baseada em valores-chave. A partir daqui, a polaridade de cada palavra é armazenada no mapa hash para processamento mais rápido. O método map () lê tweets linha a linha no arquivo e analisa cada palavra gerando tokens, onde cada token tem sua polaridade disponível no mapa hash. A polaridade é buscada para cada palavra, em seguida é calculada a polaridade geral de um único tweet usando o modelo probabilístico.

Na fase de redução, a polaridade geral de cada tweets é coletada e transformada em cinco categorias diferentes: extremamente positivos, positivos, extremamente negativos, negativos e neutros. O método reduce() funciona iterativamente para coletar vários sentimentos e com base em polaridades classifica e escreve a saída no HDFS. A figura 3 representa o fluxo por onde passam os dados até terem sua polaridade definida pelo método Naive Bayes.

**Figura 3 - Arquitetura do Sistema Proposto**




Fonte: Adaptado de Parveen e Pandey (2016)

Benevenuto et al. (2015) levantou alguns dos principais métodos capazes de definir a polaridade de sentenças e que tivessem código fonte disponível na literatura, dando um descritivo de cada um e classificando como supervisionado ou não supervisionado, conforme mostra o Quadro 1. Além disso, descreveu as saídas destes métodos e a forma como eles foram validados, conforme Quadro 2. Os códigos obtidos foram disponibilizados na página http://homepages.dcc.ufmg.br/~fabricio/benchmark_sentiment_analysis.html.

Benevenuto et al. (2015) ressalta ainda que estas abordagens ficam disponíveis para usuários que desejam aplicar a análise de sentimento, de modo que cada um escolhe alguma das soluções disponíveis e aplica em seu propósito específico. Pode-se notar, no caso dos métodos supervisionados, o uso de modelos previamente treinados com a base de dados original, em vez de após uma nova fase de treinamento.

**Quadro 1 - Métodos para Análise de Sentimentos em Sentenças**

| Nome | Descrição | TNS | TS |
|---|---|---|---|
| Emoticons | Classificação do texto de acordo com uma lista de emoticons codificada em positivos (":)") e negativos (":("). | X | |
| ANEW | Conforme descrito por Bradley e Lang (1999), Affective Norms for English Words, possui 1034 palavras em inglês. As palavras neste método foram ranqueadas com relação ao prazer, excitação e dominância, onde cada palavra é atribuída a um valor de 1 a 9. | X | |
| Happiness Index | Escala de sentimentos que utiliza o ANEW (um conjunto de palavras ligadas a emoções do Inglês), avaliando textos com valores de 1 a 0 para marcar a quantidade de felicidade existente, conforme Dodds e Danforth (2009). | X | |
| SentiWordNet | Segundo Esuli e Senastiani (2006), esse método é baseado em um léxico já conhecido, chamado WordNet, em que os autores agruparam adjetivos, substantivos e verbos em conjuntos de palavras similares para formar uma rede de palavras. No SentiWordNet é associada uma polaridade entre algumas palavras-raiz do WordNet e se propaga essa polaridade nas palavras similares da WordNet criando um amplo léxico de sentimentos. | X | X |
| SASA | Conforme Wang et. al. (2012), foi utilizado, nos Estados Unidos, para detectar sentimentos no Twitter durante as eleições presidenciais em 2012. Utiliza modelos estatísticos do classificador Naïve Bayes. | | X |
| PANAS-t | Gonçalves et al. (2013) descreve que este método é um léxico modificado do PANAS (Positive Affect Negative Affect Scale), que possui uma escala psicométrica com um grande conjunto de palavras associadas a 11 diferentes tipos de humor (surpresa, medo, etc). | X | |
| SentiStrength | Segundo Thelwall (2013), foi construído a partir de um dicionário léxico com anotações imputadas por humanos e melhorada por eles com o uso do aprendizado de máquinas. | X | X |
| AFINN | Conforme Nielsen, 2011 feito a partir do ANEW, com a diferença de ter como foco as redes sociais e por isso possuindo gírias, acrônimos e palavrões da língua Inglesa. São 2.477 termos classificados entre -5 (mais negativo) e +5 (mais positivo). | X | |
| OpinionLexion | Este método é conhecido por Sentiment Lexicon e, conforme Hu e Li (2004), é formado por uma lista com aproximadamente 6.800 palavras rotuladas como positivas e 6.800 negativas, também contém gírias e abreviações no idioma Inglês e foi criado a partir de textos de reviews de produtos em sites de compra. | X | |
| Umigon | Léxicos propostos para detectar sentimentos e subjetividade no Twitter. Ele utiliza, segundo Levallois (2013) diversos recursos linguísticos (onomatopeias, exclamações, emoticons, etc) e possui heurísticas responsáveis por desambiguar o texto baseado em negações, palavras alongadas e hashtags. | X | |
| Vader | Segundo Hutto e Gilbert (2014), possui um dicionário léxico estabelecido como LIWC, ANEW e GI, com adições de construções léxicas presentes em microblogs (emoticons, acrônimos e gírias que expressam sentimentos). | X | |

Fonte: Adaptado de Benevenuto et al. (2015)

**Quadro 2 – Saída e Validação dos Métodos de Análise de Sentimentos em Sentenças**

| Nome | Saída (Força do Sentimento) | Validação |
|---|---|---|
| Emoticons | -1,1 | - |
| ANEW | **1, 2, 3, 4, 5, 6, 7, 8, 9** | - |
| Happiness Index | **1, 2, 3, 4, 5, 6, 7, 8, 9** | Letras de músicas, Blogs, Mensagens oficiais do governo. |
| SentiWordNet | **-1,0,1** | - |
| SASA | **Negative, Neutral, Unsure, Positive** | Tweets "Políticos" rotulados por "turkers" (AMT). |
| PANAS-t | **-1,0,1** | Validação com dataset não rotulado do twitter |
| SentiStrength | **-1,0,1** | Seus próprios datasets - Twitter, Youtube, Digg, Myspace, BBC Forums and Runners World. |
| AFINN | **-1,0,1** | Twiter |
| Opinion Lexicon | **-1,0,1** | Reviews de produtos da Amazon e CNET |
| Umigon | **Negative, Neutral, Positive** | Twitter e SMS |
| Vader | **-1,0,1** | Seus próprios datasets - Twitter, Reviews de Filmes, Reviews Técnicos de Produtos, Opiniões de usuários do NYT. |

Fonte: Adaptado de Benevenuto et al. (2015)

## 3.6 Aplicações Práticas dos Métodos

Jianqiang e Xiaolin (2017) discutem os efeitos do método de pré-processamento de texto no desempenho de classificação de sentimento, avaliando os efeitos de vários métodos de pré-processamento sobre a classificação do sentimento, incluindo a remoção de URLs, a redução de letras repetidas, a remoção de palavras de parada, a remoção de números e a ampliação de siglas. Eles utilizaram dois modelos de recursos e quatro classificadores para identificar a polaridade do sentimento do Twitter em cinco conjuntos de dados do Twitter. Os resultados experimentais mostraram que o desempenho da classificação do sentimento melhora após expansão das siglas e substituição de negação, mas pouco muda ao remover URLs, palavras de parada ou números.

Saif et al. (2014) estudou o efeito de diferentes métodos de remoção de palavras de parada para a classificação de polaridade dos tweets e se a remoção de palavras de parada afeta o desempenho dos classificadores do sentimento do Twitter. Eles aplicaram seis diferentes métodos de identificação de palavras para seis conjuntos de dados diferentes do Twitter e observaram que a remoção de palavras de

parada afeta dois métodos de classificação de sentimento supervisionados. Saif et al. (2012) descobriram que o pré-processamento levou a uma redução significante do espaço de recursos. Após o pré-processamento, o tamanho do vocabulário foi reduzido em 62%.

Bao et al. (2015) explorou o efeito dos métodos de pré-processamento na classificação de sentimentos do Twitter. Eles avaliaram os efeitos de URLs, negação e letras repetidas e, diferente do que foi identificado por Jianqiang e Xiaolin (2017), os resultados experimentais no conjunto de dados do Stanford Twitter Sentiment mostraram que a precisão da classificação de sentimento aumenta quando há retirada de URL, transformação de negação e normalização de letras repetidas.

Benevenuto; Ribeiro e Araújo (2013) fizeram um trabalho usando seis eventos e duas bases de dados diferentes provenientes do Twitter, com um intuito de comparar oito métodos propostos na literatura: LIWC, Happiness Index, SentiWordNet, SASA, PANAS-t, Emoticons, SenticNet e SentiStrength. Uma das bases continha por volta de 1,8 bilhões de mensagens coletadas do Twitter, onde deste total, foram filtrados tweets associados a seis eventos sociais relacionados a tragédias, lançamento de produtos, política, saúde e esporte. A outra base de dados possuía uma coleção de textos rotulados manualmente como positivos e negativos. A partir de bases de dados reais, eles compararam os oito métodos de análise de sentimentos em termos de abrangência (fração de mensagens capturadas por cada método) e concordância (fração de sentimentos corretamente identificados por cada método).

Entre os resultados encontrados, o artigo de Benevenuto; Ribeiro e Araújo (2015) sumarizou, em quatro principais:

1. Os métodos possuem diferentes graus de abrangência, variando entre 4% e 95% quando aplicados a dados associados a eventos reais. Isso sugere que, dependendo do método utilizado, apenas uma pequena fração de mensagens será analisada, podendo levar a resultados enviesados ou não representativos.
2. Nenhum método alcançou níveis altos de abrangência e concordância ao mesmo tempo. O método Emoticons atingiu a maior acurácia (acima de 85%), porém uma das menores abrangências (4–13%).
3. A concordância dos métodos, quando aplicados aos dados rotulados, variaram entre 33% e 80%, sugerindo que uma mesma amostra de dados pode ser interpretada de forma diferente dependendo do método escolhido.

4. Existe desacordo entre os métodos na predição de sentimentos para diferentes eventos considerados. Para o caso do evento da queda de um avião, metade dos métodos detectaram mais positividade do que negatividade. O mesmo é observado em outros eventos onde eram esperados uma maior quantidade de sentimentos negativos.

Benevenuto; Ribeiro e Araújo (2015) citou que a maioria dos métodos tende a apresentar mais sentimentos positivos que negativos, já que não foram encontradas muitas curvas abaixo do “gabarito” para a base. Além disso, notaram que muitos métodos obtiveram somente valores positivos, independente da base analisada. O artigo deu destaque para o fato de o método SenticNet ter apresentado as maiores taxas de abrangência, mas ter identificado polaridades incorretas para conjuntos de dados onde predominavam as negativas.

Segundo Benevenuto; Ribeiro e Araújo (2015), o fato da maioria dos métodos tenderem a identificar os sentimentos positivos pode atrapalhar a detecção em tempo real de ferramentas desenvolvidas para esse contexto, dado que elas apenas aplicam os métodos nos dados e avaliam a percentual de mensagens positivas e negativas. É importante ressaltar que os números mostram que não existe um método que sempre obtém a melhor predição para datasets diferentes, conforme Benevenuto; Ribeiro e Araújo (2015) colocou, *“uma investigação preliminar deve ser conduzida quando se utilizar um novo dataset.”.*

# 4 ANÁLISE CRÍTICA

Sabendo que dados semiestruturados não se encaixam facilmente em bancos de dados relacionais e que a análise de sentimento de Big Data demanda altas capacidades de processamento e extravagante habilidade analítica, estudaram-se aqui ferramentas de armazenamento e processamento paralelo, tais como MapReduce e Hadoop para a exploração e execução de análises de dados coletados do Twitter

De modo geral, o que foi proposto na primeira parte deste trabalho, se trata não só de coletar os dados, mas também de armazená-los no Hadoop, uma vez que o MapReduce é uma ferramenta que otimiza de forma significante este trabalho: o Map rotula os tweets de acordo com os dados de treinamento (ou dicionário léxico) e enquadra na categoria adequada. Em seguida, o Reduce resume todas as instâncias das palavras de cada categoria e faz a contagem. O Map-Reduce, portanto, lida com a formação de modelos para os classificadores. Em suma, como visto no decorrer do trabalho, a estrutura do Hadoop é capaz de rodar em clusters de computadores e executar análises estatísticas completas para enormes quantidades de dados.

A Análise de Sentimentos do Twitter oferece às organizações capacidade de monitorar o sentimento público em relação aos produtos e eventos relacionados a eles em tempo real. Para tal, após a coleta dos dados, um importante passo antes da análise, é o pré-processamento. Na seção 4.1, apresentou-se a importância de efetuar este primeiro passo e algumas consequências de não fazê-lo, dado o contexto dos tweets, que geralmente são compostos por frases incompletas, ruidosas e mal estruturados, expressões irregulares, palavras mal formadas e termos que não estão no dicionário. Antes da seleção de recursos, uma série de pré-processamentos (por exemplo, remoção de palavras de parada e URLs) são aplicadas para reduzir a quantidade de ruído nos tweets. Destaca-se a seguir, analisando experimentos apresentados no capítulo 3.6, como a aplicação de distintos métodos de pré-processamento podem elevar o desempenho da análise do sentimento.

- Em alguns métodos a remoção de URL’s muda de forma significante os resultados, em outros não há impacto. Isso significa que antes de efetuar a análise, necessita-se verificar se o método que será utilizado é prejudicado com a existência das URL’s, caso não, não há necessidade desse esforço;

- Remoção de números não tem efeito sobre a precisão da classificação de sentimento no modelo de polaridade anterior porque os números são neutros.
- O efeito de remover letras repetidas no desempenho de classificadores é diferente em cada conjunto de dados, o que sugere que a remoção de letras repetidas influencia a polaridade e as características semânticas das palavras nos tweets.
- A ampliação das siglas melhora o desempenho dos classificadores na maioria dos conjuntos de dados e métodos;
- O desempenho dos classificadores aumenta após a substituição da negação em todos os conjuntos de dados, na maioria dos casos porque a negação contém características importantes de polaridade do sentimento;
- O desempenho do SVM aumenta depois de expandir as siglas e substituir as negações, portanto, estes são métodos de pré-processamento efetivos ao usar o classificador SVM;
- A deleção aleatória de palavras de parada causa um declínio significativo no desempenho da classificação porque a palavra excluída aleatoriamente pode ser uma palavra-chave faltando, causando a polaridade ou o dano da decisão da relação semântica de sentença. Por exemplo, na frase "*Eu não gosto de celulares grandes*", com a remoção das palavras de parada restaria apenas "*gosto celulares grandes*", o que altera completamente o sentimento contido na frase.

Além dos itens citados, identifica-se que o método de pré-processamento afeta o desempenho dos classificadores de sentimentos de forma semelhante, enquanto os classificadores NB e RF são mais sensíveis do que os classificadores LR e SVM. Um fator que pode afetar os resultados da classificação do sentimento é a escolha do classificador e os recursos usados para o treinamento.

Diante destas constatações, nota-se a importância do pré-processamento, entretanto, é necessário tomar cuidado em onde, qual tipo e em quais casos cada método deve ser utilizado, por exemplo, quando se realiza o pré-processamento sem considerar emoticons, os tweets que contém sentimentos na forma de emoticons podem ser simplesmente ignorados pelos algoritmos, portanto, quando trata-se de um conjunto de dados que possui um quantidade considerável de emoticons, considera-los no pré-processamento deixa os resultados mais precisos.

Outro exemplo em que deve-se analisar antes de aplicar o pré-processamento é na retirada de letras repetidas em palavras, dado que, muitas vezes elas estão desta

forma pois o usuário ao escrever quis dar ênfase na mesma, o que pode ser valioso na análise. Desde modo, retirar letras repetidas das palavras no pré-processamento exige o cuidado de substituir três ou mais por duas, de modo que seja possível distinguir quando uma palavra está sendo enfatizada ou não.

Diante de todos os métodos apresentados nos capítulos 3.5 e 3.6 e sabendo que cada método possui suas particularidades e pode ser mais bem aproveitado em distintas situações é possível fazer uma análise para identificar quando determinados procedimentos fazem sentido em cada caso.

O método de Emoticon, como intuitivamente o nome diz, se aplica de forma mais assertiva e contextos que possuem uma grande quantidade de emoticons no conjunto de dados e efetuando o pré-processamento considerando-os. Já o método ANEW pode ser bem aproveitado quando se deseja coletar dados relacionados a torcidas esportivas ou quaisquer situações onde haja agitação e euforia por partes de quem está postando, dado que este método ranqueia as palavras com relação a excitação e dominância.

O método SASA é comprovadamente muito eficiente para análise de sentimentos de textos relacionados à política, dado que já foi utilizado neste contexto. Quando se deseja apenas detectar flutuações de humor dos usuários, o método recomendado é o PANAS-t.

O método AFINN, que foi desenvolvido a partir do ANEW, também é indicado para casos de torcidas esportivas, entretanto, ele possui gírias, siglas e palavrões o que o torna mais adequado para o contexto do Twitter. O método OpinionLexion também contém gírias e abreviações, entretanto, ele se diferencia pelo fato de ter sido criado a partir de textos de produtos em sites de compra, sendo assim, entende-se que ele pode ser adequado caso o desejo seja avaliar o sentimento de tweets relacionados a um determinado produto (a opinião dos usuários com relação a um smartphone, por exemplo).

Já o Naive Bayes, um dos métodos mais utilizados no contexto de análise de sentimentos (dado que é um dos métodos de aprendizagem mais práticos) é recomendado quando há grandes conjuntos de treinamento disponíveis e os atributos que expõem as instâncias forem independentes, isto é, quando se deseja analisar as sentenças considerando que uma palavra contém sua polaridade independente dos demais termos da sentença, uma vez que é desta forma que este método funciona.

Quando se fala em SVM, pode-se dizer que é um método decisivo para fatos em que é indispensável um alto poder de previsão. Eles são mais difíceis de visualizar do que o NB, por exemplo, devido à complexidade na formulação, entretanto, são utilizadas principalmente quando o objetivo for encontrar padrões ou tendências que auxiliem no entendimento dos dados.

# 5 CONSIDERAÇÕES FINAIS

Este trabalho enfatizou a importâcia da mineração de dados oriundos das Redes Sociais e o interesse das empresas, marcas, políticos e artistas na análise destes dados. Ao exibir algumas funcionalidades e características do Twitter, mostrou-se que o fato de existir um limite baixo de caracteres é um dos fatores que o difere das demais RSO, atraindo os usuários a expressarem suas opiniões curtas em tempo real a respeito de um determinado tema.

Sabendo a quantidade de dados gerados a cada instante nestas redes, notou-se que os Bancos de Dados tradicionais não são mais suficientes para tratar de forma eficiente desses dados, introduzindo, assim, os conceitos de Big Data. Apresentaram-se métodos de coleta e armazenamento sendo aplicados na prática e foram apresentadas algumas ferramentas do Ecossistema Hadoop, mostrando técnicas utilizadas na tarefa de ingestão de Big Data.

Foram apresentados os procedimentos de obtenção da polaridade dos sentimentos dos tweet, passando por métodos de pré-processamento e análise. Por fim, foram apresentadas as vantagens e desvantagens de alguns métodos com relação a outros em determinadas situações, evidenciando comparações entre os procedimentos de análise antes e após o pré-processamento, com o intuito de mostrar a importância desta tarefa em meio ao processo como um todo.

Houve uma certa dificuldade de encontrar trabalhos que passassem por todos os procedimentos da Análise de Sentimentos, desde a coleta dos dados até a análise propriamente dita. Notou-se que não há um método, algoritmo ou procedimento que possa ser dito como melhor ou mais eficiente para todos os casos. Diversos trabalhos citados neste efetuaram os procedimentos de análise, aplicando diferentes métodos aos mesmos conjuntos de dados e obtendo diferentes resultados, o que possibilitou evidenciar que para cada situação, conjunto de dados ou mesmo objetivo, tem-se um método que se adequa melhor. Deste modo, antes de efetuar o trabalho de análise é necessário identificar as características do dataset, o público alvo e o objetivo que é pretendido alcançar.

# REFERÊNCIAS

AGARWAL, A. et. al. Sentiment Analysis of Twitter Data. Department of Computer Science Columbia University: New York, 2011.

APACHE FLUME. Welcome to Apache Flume. 2017. Disponível em: <http://flume.apache.org/> Acesso em: 17 de fevereiro de 2018.

APACHE HADOOP. Welcome to Apache Hadoop. 2017. Disponível em: <http://hadoop.apache.org/> Acesso em: 17 de fevereiro de 2018.

APACHE HADOOP. HDFS Architecture Guide. 2013. Disponível em: <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html> Acesso em: 17 de fevereiro de 2018.

APACHE SQOOP. The Apache Software Fundation. 2018. Disponível em: <http://sqoop.apache.org/> Acesso em: 17 de fevereiro de 2018.

ARAUJO, M.; GONÇALVES, P.; BENEVENUTO, F. Métodos para Análise de Sentimentos no Twitter. In: Proceedings of the Simposio Brasileiro de Sistemas Multimídia e Web, 2013.

BENEVENUTO, F. et. al. Métodos para Análise de Sentimentos em Mídias Sociais. In: Proceedings of the Simposio Brasileiro de Sistemas Multimídia e Web, 2015.

BRADLEY, M. M.; LANG, P. J. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, 1999.

CAMBRIA, E. et. al. Senticnet: A publicly available semantic resource for opinion mining. In AAAI Fall Symposium Series, 2010.

CARVALHO, J. A. F. Mineração De Textos: Análise De Sentimento Utilizando Tweets Referentes À Copa Do Mundo 2014. QUIXADÁ, 2014.

CHEN, H. Knowledge management systems: a text mining perspective. Arizona: Knowledge Computing Corporation, 2001.

COSTA, L. H. M. K. et. al. Grandes Massas de Dados na Nuvem - Desafios e Técnicas para Inovação, 2012.

DODDS, P. S.; DANFORTH, C. M. Measuring the happiness of large-scale written expression: songs, blogs, and presidents, 2009.

EFSTATHIADES, H. et. al. Distributed Large-Scale Data Collection in Online Social Networks. Department of Computer Science, University of Cyprus, 2016.

ESULI, A.; SEBASTIANI, F. Sentiwordnet: A publicly available lexical resource for opinion mining, 2006.

FRANÇA, C. T. et al. Big Social Data: Princípios sobre Coleta, Tratamento e Análise de Dados Sociais. SBC: 2014. 1a ed.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classication using distant supervision, Stanford, 2009.

GONÇALVES, C. A. Análise de Sentimentos em Reclamações. Rio de Janeiro, 2016.

GONÇALVES, P. et. al. Comparing and combining sentiment analysis methods. In Proc. COSN, 2013.

HARRINGTON, P. Machine learning in action, Manning Publications Co., 2012.

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

HUE. How-to: Analyze Twitter Data with Hue. 2013. Disponível em: <http://gethue.com/how-to-analyze-twitter-data-with-hue/> Acesso em: 17 de fevereiro de 2018.

HUTTO, C. J.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In ICWSM, 2014.

JIANQIANG, Z.; XIAOLIN, G. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. IEEE Access, v. 5, pp. 2870-2879, 2017.

JIANQIANG, Z. Pre-processing boosting Twitter sentiment analysis? Smart City, pp. 748-753, 2015.

JINTURKAR, M.; GOTMARE, P. Sentiment Analysis of Customer Review Data using Big Data: A Survey. Emerging Trends In Computing, 2016.

LEVALLOIS, C. Umigon: sentiment analysis for tweets based on terms lists and heuristics. Atlanta, Georgia, USA. Association for Computational Linguistics, 2013.

LI, Y. M.; LI, T. Y. Deriving marketing intelligence over microblogs. In: Proceedings of 44th Hawaii International Conference on System Sciences (HICSS), p. 1 –10, 2011.

LUNARDI, A. C.; VITERBO, J.; BERNARDINI, F. C. Um Levantamento do Uso de Algoritmos de Aprendizado Supervisionado em Mineração de Opiniões. Rio de Janeiro, 2016.

NARR, S.; HULFENHAUS, M.; ALBAYRAK, S. Language-independent Twitter sentiment analysis, Knowledge Discovery and Machine Learning, pp. 12-14, 2012.

NASCIMENTO, P.; OSIEK, A. B.; XEXEO, G. ANÁLISE DE SENTIMENTO DE TWEETS COM FOCO EM NOTÍCIAS. Revista Eletrônica de Sistemas de Informação: 2015. v. 14, n. 2.

NIELSEN, F. A. A new anew: Evaluation of a word list for sentiment analysis in microblogs, 2011.

PARVEEN, H.; PANDEY, S. Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm, 2016.

SAIF, H. et. al. On stopwords, ltering and data sparsity for sentiment analysis of Twitter. Reykjavik, Iceland, pp. 80-81, 2014.

SAIF, H. et. al. Alleviating data sparsity for Twitter sentiment analysis. CEUR Workshop, pp. 2-9, 2012.

SAIF, H.; FERN, M.; HE, Y. Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-gold. Italy, pp. 21-26, 2013.

STELZNER, M. A. How Marketers Are Using Social Media to Grow Their Businesses. SOCIAL MEDIA MARKETING INDUSTRY REPORT: 2012.

SUPRAJA, G.S.; MUJAMDAR, J.; ANKALAKI, S. A Big Data Methodology for Sentiment Analysis of Twitter Data. International Journal of Innovative Research in Computer and Communication Engineering, 2015

TAUSCZIK, Y. R; PENNEBAKER, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, Journal of Language and Social Psychology, v. 29, n. 1, pp. 24-54, 2009.

THELWALL, M. Heart and soul: Sentiment strength detection in the social web with sentistrength, 2013.

THELWALL, M.; BUCKLEY, K.; PALTOGLOU, G. Sentiment strength detection for the social Web, J. Amer. Soc. Inf. Sci. Technol., v. 63, no. 1, pp. 163-173, 2012.

VOHRA, S. M.; TERAIYA, J. B. A Comparative Study Of Sentiment Analysis Techniques. Journal Of Information, Knowledge And Research In Computer Engineering: v. 02, 2013.

WANG, H. et. al. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In ACL System Demonstrations, 2012.

ZHAO, D.; ROSSON, M. B. How and why people twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work, p. 243-252, 2009.

# APÊNDICES

## *Apêndice I*

1) INICIAR O SERVIÇO DA CLOUDERA
/home/training/scripts/analyst/da_toggle_services.sh

2) IR PARA O DIRETORIO DO FLUME
cd /usr/lib/flume-ng/lib

3) CRIAR O ARQUIVO “/home/training/twitter.conf” CONTENDO:

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = org.flume.source.twitter.json.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = chave gerada pelo twitter
TwitterAgent.sources.Twitter.consumerSecret = chave gerada pelo twitter
TwitterAgent.sources.Twitter.accessToken chave gerada pelo twitter
TwitterAgent.sources.Twitter.accessTokenSecret = chave gerada pelo twitter
TwitterAgent.sources.Twitter.keywords = termo que deseja buscar nos tweets
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = /user/flume/tweets
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 1000
```

4) SUBIR O SERVIÇO DO FLUME
“flume-ng agent -f /home/training/twitter.conf Dflume.root.logger=DEBUG,console -n TwitterAgent”

5) CRIAÇÃO DA TABELA:

```
CREATE TABLE twitter (
json string
)
LOCATION '/user/flume/tweets/';
```

6) SELEÇÃO DOS CAMPOS:

```
SELECT
get_json_object(json, '$.id') AS ID,
get_json_object(json, '$.created_at') as CREATED_AT,
get_json_object(json, '$.text') as TEXT,
get_json_object(json, '$.user.name') as NAME,
get_json_object(json, '$.user.location') as LOCATION,
get_json_object(json, '$.user.screen_name') as SCREEN_NAME
FROM twitter;
```

# ANEXOS

## *Anexo I*

CÓDIGO USANDO STREAMING API

```
import json
from requests_oauthlib import OAuth1Session
key = "{sua API key}"
secret = "{sua API secret}"
token = "{seu Acess Token}"
token_secret = "{seu Acess Token Secret}"
requests = OAuth1Session(key, secret, token, token_secret)
r = requests.post('https://stream.twitter.com/1/statuses/filter.json',
data={'track': 'bom dia'},
stream=True)
for line in r.iter_lines():
if line:
print json.loads(line) # tweet retornado
```

## *Anexo II*

CÓDIGO USANDO REST API

```
import oauth2 as oauth
import json
import time
CONSUMER_KEY = "{sua API key}"
CONSUMER_SECRET = "{sua API secret}"
ACCESS_KEY = "{seu Acess Token}"
ACCESS_SECRET = "{seu Acess Token Secret}"
consumer = oauth.Consumer(key=CONSUMER_KEY,
secret=CONSUMER_SECRET)
access_token = oauth.Token(key=ACCESS_KEY, secret=ACCESS_SECRET)
client = oauth.Client(consumer, access_token)
q = 'israel' #termo a ser buscado
url =
"https://api.twitter.com/1.1/search/tweets.json?q="+str(q)+"&count=100"+"&lang=pt"
response, data = client.request(URL, "GET")
tweets = json.loads(data)
for tweet in tweets['statuses']:
print str(tweet)
```