Gabriel Silva Monteles

ANÁLISE DE SENTIMENTO: uma comparação de dados extraídos do Twitter a partir de diferentes dicionários léxicos

São Luís - MA

2019

Gabriel Silva Monteles

ANÁLISE DE SENTIMENTO: uma comparação de dados extraídos do Twitter a partir de diferentes dicionários léxicos

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Orientadora: Prof.ª Dra. Simara Vieira da Rocha

São Luís - MA

2019

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Monteles, Gabriel Silva.
ANÁLISE DE SENTIMENTO: uma comparação de dados extraídos do Twitter a partir de diferentes dicionários léxicos / Gabriel Silva Monteles. - 2019.
54 f.

Orientador(a): Simara Vieira da Rocha.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, Universidade Federal do Maranhão, São Luís, 2019.

1. Análise de Sentimentos. 2. Big Data. 3. Dicionários Léxicos. I. Rocha, Simara Vieira da. II. Título.

Gabriel Silva Monteles

ANÁLISE DE SENTIMENTO: uma comparação de dados extraídos do Twitter a partir de diferentes dicionários léxicos

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do grau de bacharel em Ciência da Computação.

Aprovado em:

Prof.ª Dra. Simara Vieira da Rocha
Orientadora
Universidade Federal do Maranhão

Prof. M.Sc. Carlos Eduardo Portela Serra de Castro
Examinador
Universidade Federal do Maranhão

Prof. Dr. Ivo José da Cunha Serra
Examinador
Universidade Federal do Maranhão

São Luís - MA
2019

Agradecimentos

À minha família, por ter me apoiado em todos os momentos, pelas repreensões e ensinamentos que contribuíram para tornar a pessoa que sou hoje.

Aos meus pais, pelo incentivo, cuidado e amor e carinho que me concederam, além de todos os esforços que tiveram para me dar tudo do melhor jeito possível.

Aos meus avós, que sempre cuidaram de mim e sempre lutaram e torceram para o meu sucesso.

À minha irmã, Gabrielle, que me auxiliou na vida e nos estudos, desde o início, para que eu me tornasse o que sou hoje, como estudante e ser humano.

À minha namorada, Ednara, por estar comigo nos últimos anos da faculdade, me apoiando e me ajudando em todas as escolhas tomadas, me fazendo evoluir muito como ser humano.

Aos meus amigos Caio, Elias, Frank e Pacheco, por estarem comigo desde o ensino médio, sempre torcendo para o meu sucesso.

À minha orientadora, Prof.ª Dra. Simara, por aceitar ser minha orientadora e ter tido paciência, competência e dedicação comigo.

Ao Professor Geraldo e ao PETComp, por terem me proporcionado desafios, experiências e viagens durante o curso, que ajudaram a moldar o meu caráter e a me tornar a pessoa e profissional que sou hoje.

Aos True Friends, por estarem comigo em muitos momentos do curso, da vida e nas viagens, proporcionando momentos únicos na minha vida.

Aos amigos do CodeBuilders, que desde o início do curso enfrentam desafios e realizam conquistas comigo, além de manter unida a turma de 2015.1 e amigos do curso.

À Atlética Lorde, onde criei grandes amizades e vivi momentos especiais, além de ter criado grande apego e espirito de coletividade dentro do curso.

Ao time de vôlei da Atlética Lorde, responsável por me mostrar novos hábitos, me dar muitas alegrias durante o curso e durante vários momentos juntos nas competições e na vida.

À todos aqueles que contribuíram diretamente e indiretamente para a realização deste trabalho.

"Persistence is the shortest path to success!" (Charles Chaplin)

Resumo

Ao longo dos últimos anos, o número de dispositivos conectados a internet cresceu exponencialmente e, consequentemente, houve um aumento da quantidade de dados gerados por esses dispositivos. Essa grande quantidade de dados gerados a todo momento, deu origem ao conceito de *Big Data*. Esse volume de dados possibilita a prática da análise de sentimentos, onde é possível classificar opiniões em positivas, negativas ou neutras em diversos meios, como em *tweets*, onde, por apresentar comentários com poucos caracteres, torna-se viável a produção de uma análise a nível de sentença. Um dos principais problemas relacionados a análise de sentimentos a nível de sentença está relacionado à declarações onde o sentimento de uma determinada sentença não permanece explicito. Também consistem em problemas, sentenças onde é utilizada uma linguagem informal ou são utilizados gírias e regionalismos, impossibilitando a classificação da polaridade de determinada palavra, acarretando em uma classificação errônea de determinado texto. Almejando uma análise que possa contornar os problemas apresentados, neste trabalho será realizada uma análise de sentimentos a nível de sentença tendo como base *tweets* a respeito da grande quantidade de focos de incêndio ocorridos na região amazônica no mês de setembro de 2019. Essa análise será feita visando uma comparação entre os três dicionários léxicos utilizados na metodologia, onde será proposto um conjunto de melhorias nos mesmos, objetivando a eliminação de inconsistências que acarretam em problemas comuns durante a análise de sentimento a nível de sentença e, consequentemente, produzir uma análise com resultados mais precisos. Por fim, os resultados obtidos pelos dicionários após as modificações foram excelentes, acarretando em um resultado das avaliações dos dicionários semelhante à avaliação realizada pelo autor.

Palavras-chaves: *Big Data*. Análise de Sentimentos. Twitter. Dicionário Léxico.

Abstract

Over the last few years, the number of devices connected to the internet has grown exponentially and, consequently, there has been an increase in the amount of data generated by these devices. This large amount of data generated at all times gave rise to the concept of Big Data. This volume of data enables the practice of sentiment analysis, where it is possible to classify opinions as positive, negative or neutral in various media, such as in Tweets, where, by presenting comments with few characters, it is feasible to produce a sentence-level analysis. One of the main problems related to sentence-level sentiment analysis is related to statements where the sentiment of a particular sentence does not remain explicit. There are also problems, sentences in which informal language is used or slang and regionalisms are used, making it impossible to classify the polarity of a given word, resulting in a misclassification of a given text. Aiming at an analysis that can circumvent the problems presented, this paper will analyze sentiments at the sentence level based on the large amount of fire outbreaks that occurred in the Amazon region in September 2019. This analysis will be made aiming at a comparison between the three Lexical dictionaries used in the methodology, where a set of improvements will be proposed, aiming at the elimination of inconsistencies that lead to common problems during the sentence-level disagreement analysis and, consequently, produce an analysis with more accurate results. Finally, the results obtained by the dictionaries after the modifications were satisfactory, resulting in a result of the dictionary evaluations similar to the author's evaluation.

Keywords: *Big Data*. Sentiment Analysis. Twitter. Lexicon Dictionary.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	Application Programming Interface
GB	Gigabyte
GFS	Google File System
HD	Hard Disk
HDFS	Hadoop File System
HTTP	Hypertext Transfer Protocol
IDE	Integrated Development Environment
IMDG	In-Memory Data Grid
JSON	JavaScript Object Notation
JVM	Java Virtual Machine
LIWC	Linguistic Inquiry and Word Count
NoSQL	Not Only SQL
PHP	Hypertext Preprocessor
RAM	Random Access Memory
RDBMS	Relational Database Management Systems
SQL	Structured Query Language
SVM	Support Vector Machine
TB	Terabyte

Sumário

1 Introdução

Nos últimos anos, o crescimento e surgimento de novas redes sociais se intensificaram no Brasil e no mundo. Cerca de 1 milhão de pessoas ganham acesso à internet por dia no mundo e, consequentemente, o número de usuários de redes sociais aumenta devido a isso. Além disso, dos 10 sites mais acessados no mundo, quatro são redes sociais, entre elas está o Twitter (REVISTA PLANETA, 2019a). Em consequência a isso, não só o Twitter, mas também outras redes sociais, contribuem significativamente com uma grande quantidade de dados gerados na internet.

Companhias estão aproveitando o grande volume de dados para construir informações úteis em diversas áreas. Esse movimento, conhecido como a era do *Big Data*, está ganhando impulso nos últimos anos, movido pelo crescimento do poder computacional e do surgimento de novas fontes geradoras de informação, como dados de dispositivos móveis e das redes sociais (CHEN et al., 2014).

Para Beyer e Laney (2012), *Big Data* é definido como grande volume e/ou alta variedade de informações importantes que requerem novas formas de processamento que permitam capacidade de decisão apurada, descobertas de *insight* e otimização de processamento.

Para Liu (2012), a análise de sentimentos é uma área de pesquisa que investiga as opiniões das pessoas para diferentes materiais: produtos, eventos e organizações. Com o *Big Data* em expansão, a análise de sentimentos se torna pertinente na atualidade, visto que a grande quantidade de dados contribuem para uma maior possibilidade de uso para os mesmos.

Com o intuito de aproveitar os dados gerados pelas redes sociais, grandes empresas fazem uso da análise de sentimentos (ou mineração de opinião) para que possam obter um *feedback* de seus clientes sobre os seus produtos, assim, obtendo um suporte durante o processo de tomada de decisão. Essas opiniões podem ser utilizadas para tomada de decisão em diversas áreas de atuação. Uma dessas áreas de atuação é a do comércio eletrônico, onde o *feedback* de diferentes usuários pode ajudar na classificação do impacto de um determinado produto lançado no mercado (REVISTA PLANETA, 2019b).

Na área da saúde, Ruiz (2016) produziu um trabalho em que análise de sentimentos foi realizada, também, em cima de *tweets* relacionados a campanhas publicitárias de saúde do governo, visando o impacto dessas campanhas diante do público. Além disso, o autor ressalta a importância do fenômeno *Health self-reporting* (“auto-relato de saúde”, em tradução livre), em que usuários relatam informações sobre sua saúde e, através disso, em conjunto com outros usuários, contribuem para a disponibilização de informações sobre a

saúde da população em tempo real.

A análise de sentimentos também é utilizada no mercado financeiro, onde investidores utilizam as ferramentas da análise para capturar informações de pessoas sobre determinadas empresas que possuem ações na bolsa de valores, visando a compra e venda de ações desta empresa. Do mesmo modo, na política, é possível visualizar opiniões de pessoas sobre determinados políticos em períodos de campanha eleitoral, analisando a aceitação e rejeição de determinado candidato.

Segundo Moreira et al. (2016), apesar dos diversos esforços feitos em prol da melhoria dos métodos de análise, muito ainda precisa ser realizado para que, de fato, o nível de acurácia se aproxime da forma humana de avaliar sentimentos. Grande parte das dificuldades atualmente existentes na análise de sentimentos diz respeito à linguagem dos textos avaliados, à detecção de ironia e a dificuldade de tratar a subjetividade dos textos.

A análise de sentimento a nível de sentença, que será realizada neste trabalho, é responsável por identificar sentimentos e emoções em um determinado texto. Há duas principais formas de construir uma análise de sentença a nível de sentença, assim como há diversos graus de dificuldade. Na primeira delas, são utilizados algoritmos de aprendizado de máquina. Na segunda forma, que será a utilizada, a análise é realizada a partir de técnicas baseadas em léxico, que utilizam recursos responsáveis por conter a pontuação de determinada palavra no cálculo de polaridade da mesma.

Neste trabalho, os recursos léxicos utilizados se assemelham a dicionários comuns, contendo uma palavra relacionada a uma pontuação. Por este motivo, chamaremos esses recursos de dicionários léxicos. Em relação aos desafios da análise de sentimento, é possível citar o uso de ironias e de opiniões implícitas, onde não é deixado claro o sentimento da sentença que o dono da opinião deseja transmitir e, principalmente, o uso de gírias, regionalismos e abreviações, que são bastante frequentes em redes sociais, principalmente no Twitter, onde o uso de caracteres é limitado.

A fim avaliar o desempenho de diferentes dicionários léxicos e promover o uso de diferentes ferramentas para um grande conjunto de dados, realizaremos uma análise de sentimentos a nível de sentença com a utilização de três deles, contando com o uso do *framework* Apache Ignite e o com o modelo de programação MapReduce. Os dicionários utilizados no trabalho, assim como as ferramentas, serão descritos nos capítulos seguintes.

1.1 Objetivos

1.1.1 Objetivo geral

Este trabalho tem por objetivo realizar uma análise comparativa de dados extraídos do Twitter utilizando três dicionários léxicos diferentes, como forma de analisar o possível

impacto na acurácia dos resultados produzidos pela análise de sentimento.

1.1.2 Objetivos específicos

- Analisar a eficiência de diferentes dicionários léxicos em uma abordagem léxica.
- Identificar os pontos fracos dos dicionários léxicos escolhidos e propor melhorias para os mesmos.
- Seguir uma metodologia para realização da análise de sentimento a partir de dados extraídos do Twitter utilizando os dicionários léxicos escolhidos.
- Realizar uma comparação dos resultados produzidos como forma de analisar o impacto do uso dos dicionários para a melhoria da análise de sentimentos.

1.2 Trabalhos Relacionados

No trabalho de Costa (2017), foi proposto um método para análise de sentimento a nível de sentença. Nessa metodologia, foi utilizado um dicionário léxico para que fosse determinada a polaridade de *tweets* à respeito do relançamento do console Super Nintendo, em 2017. Costa (2017) analisou manualmente 94 *tweets*, onde 62 foram classificados como positivos e 32 foram classificados como negativos. A aplicação classificou 63 *tweets* positivos e 31 como negativos, entretanto, da quantidade classificada como positivo, apenas 43 realmente eram positivos e 16 realmente eram negativos.

Visando comparar as técnicas de aprendizado de máquina e de léxico, Kolchyna et al. (2015) realizou uma análise de sentimentos a nível de sentença, também em *tweets*. No método léxico, o autor utilizou um método logaritmo no cálculo de polaridade das palavras. O autor também utilizou diferentes combinações de recursos léxicos a partir de um dicionário léxico, um dicionário de *emoticons*[1] e um dicionário gerado de maneira automática. A combinação de recursos léxicos que obteve os melhores resultados foi a que combinava o dicionário léxico comum com o dicionário de *emoticons*. O resultado obtido na melhor combinação foi de 61,74% *tweets* classificados corretamente em comparação com seus valores atribuídos manualmente pelo autor.

No trabalho de Evangelista e Padilha (2014), também foram utilizadas a abordagem léxica e a técnica de aprendizado de máquina em *posts* do Twitter e do Facebook. Em relação a análise realizada em *tweets* com abordagem léxica, o autor analisou *tweets* de três empresas brasileiras X, Y e Z, de comércio eletrônico. A análise ocorreu com a utilização do dicionário *SentiWordNet* e os resultados foram comparados com uma avaliação manual dos *tweets* realizadas pelo autor. Em relação a empresa X, o autor avaliou 21 *tweets*, obtendo

[1] Figura formada a partir da união de caracteres representando emoções

uma taxa de 57% de acerto em comparação com a avaliação manual. Dos *tweets* da empresa Y, o autor avaliou 13 *tweets* obtendo 52% de acerto. Da empresa Z, 38 publicações foram avaliadas, obtendo, também, 52% de acerto.

1.3 Organização do trabalho

Além deste capítulo introdutório, esta monografia contém outros três capítulos.

No capítulo 2, será apresentada a fundamentação teórica necessária para compreensão do trabalho. Além dos conceitos de análise de sentimentos, *Big Data* e suas tecnologias também serão fundamentadas, assim como as ferramentas utilizadas no trabalho.

No capítulo 3, a metodologia proposta no trabalho será demonstrada por meio de um estudo de caso envolvendo informações extraídas do Twitter.

Por último, no capítulo 4, as conclusões acerca deste monografia serão feitas, assim como sugestões para trabalhos futuros.

2 Fundamentação teórica

Este capítulo apresenta a fundamentação teórica das plataformas e *frameworks* utilizados no trabalho, assim como os principais métodos e tecnologias presentes no desenvolvimento do mesmo. Conceitos de *Big Data*, *frameworks* Apache Ignite e o Apache Hadoop também serão fundamentados, assim como o modelo de programção MapReduce e as definições de Análise de Sentimentos.

2.1 Big Data

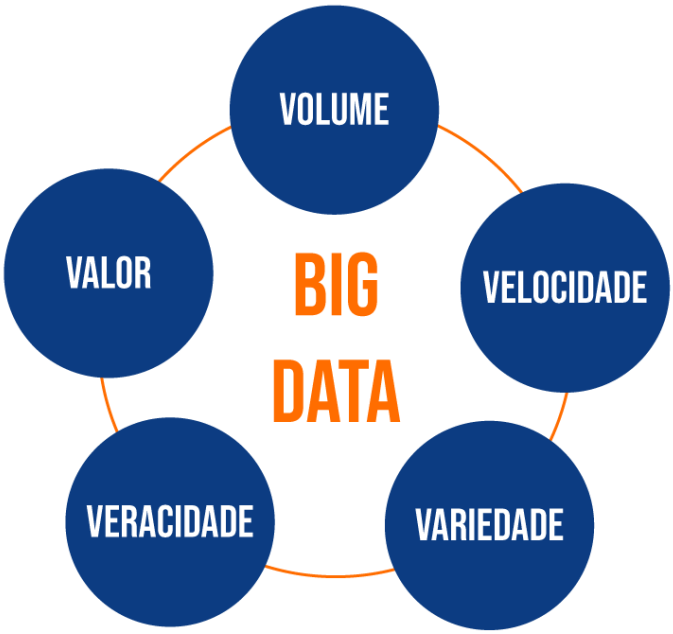
Apesar do conceito de Big Data ainda estar em discussão, segundo Chen et al. (2014), vídeos, imagens e todo tipo de informação gerada por usuários em redes sociais e quaisquer sites na internet são exemplos de fonte de dados e, para ele, *Big Data* é o conjunto de dados que crescem de forma desestruturadas e que são incapazes de serem armazenados por banco de dados convencionais.

A definição de *Big Data* consiste em um grande volume, velocidade e variedade de informações que necessitam novas formas de processamento que permitam uma melhor capacidade de processamento. Segundo Beyer e Laney (2012), essa definição de *Big Data* é conhecida como "os três Vs da *Big Data*". Além desses "três Vs", foram adicionados outros dois, veracidade e valor, que buscam representar, respectivamente, a incerteza e confiabilidade dos dados, o último ainda busca ressaltar a importância de todos os "Vs" e da necessidade de estudo de *Big Data* no cenário mundial.

2.1.1 Propriedades do Big Data

Como definido por Gadomi e Haider (2015), o *Big Data* pode ser classificado pelas seguintes propriedades: variedade, velocidade e volume. Após a reformulação do conceito, foram adicionados os outros dois valores: veracidade e valor. Na Figura 1 podem ser observadas as cinco propriedades do *Big Data* .

Figura 1 – Os Vs do *Big Data*



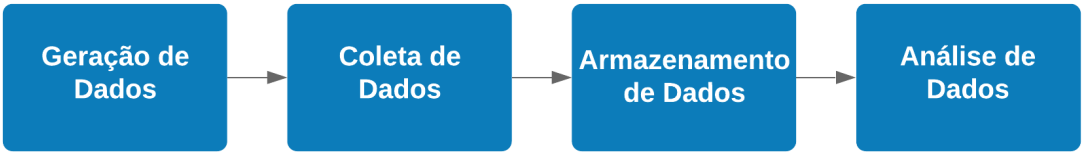
Fonte: (FUNÇÃO SISTEMAS, 2019)

- **Volume:** essa propriedade é a mais importante do *Big Data*. Ela refere-se ao grande volume de informações geradas a partir de fotos, vídeos, *e-mails* e qualquer dispositivo que produza algum tipo de informação.
- **Velocidade:** a rapidez em como os dados são transmitidos e tratados caracterizam essa propriedade do *Big Data*. Dado ao grande número de dispositivos com acesso a internet e capazes de produzir algum tipo de informação, essa característica faz-se presente no *Big Data*.
- **Variedade:** com um grande número de dados, uma grande diversidade dos mesmos são processados. Há três tipos de dados: os dados estruturados, não estruturados e semiestruturados.
- **Veracidade:** esta propriedade faz referência ao nível de confiabilidade dos dados. É necessário que a qualidade dos dados seja mantida. Ao se tratar de sentimentos, um grande desafio é lidar com a incerteza dos mesmos.
- **Valor:** uma grande quantidade de dados, sem nenhuma utilidade agregada, torna-se inútil. O quinto "V" do *Big Data* faz referência a todo esse valor que uma grande quantidade de dados deve ter para que possa ser influenciável no mercado. Uma grande volume de dados sem valor é apenas um conjunto de dados brutos.

2.1.2 Etapas do Big Data

O conjunto de etapas do *Big Data* é chamado de cadeia de valor. Essa cadeia representa os passos necessários para transformar um dado em uma informação útil no processo de tomada de decisão. De acordo com Cavanillas et al. (2016), essa cadeia de valor, que pode ser ilustrada na Figura 2, possui quatro etapas: geração de dados, coleta de dados, armazenamento de dados e análise de dados.

Figura 2 – Cadeia de valor do Big Data



Fonte: Elaborada pelo autor

- **Geração de Dados:** essa etapa corresponde ao processo de produção de informação, que pode vir de redes sociais, blogs, fóruns, smartphones e quaisquer outros dispositivos conectados à internet.
- **Coleta de Dados:** a etapa de coleta de dados é a que pode exigir um maior trabalho dentre as etapas da cadeia de valor, nela ocorre a coleta, filtragem e limpeza dos dados a serem trabalhados.
- **Armazenamento de Dados:** no armazenamento de dados, a necessidade de acesso rápido das aplicações aos dados deve ser satisfeita. Bancos de dados não relacionais e outros tipos de sistemas de arquivos são as principais alternativas para o armazenamento de um grande volume de dados sem tolerância a falhas.
- **Análise de Dados:** o principal objetivo na análise de dados é transformar os dados brutos em dados passíveis de informação, para que tenham um valor agregado e possam ser usados na tomada de decisão.

2.2 Plataformas e *Frameworks Open Source* para Análise de *Big Data*

Para Chen et al. (2014), a análise de dados é a fase mais importante na cadeia de valor de *Big Data*, pois nela originam-se as informações que irão ser entregues aos gestores para a tomada de decisão em relação a atividade exercida. Uma plataforma *open source* compreende em uma ferramenta ou programa que possui o código fonte

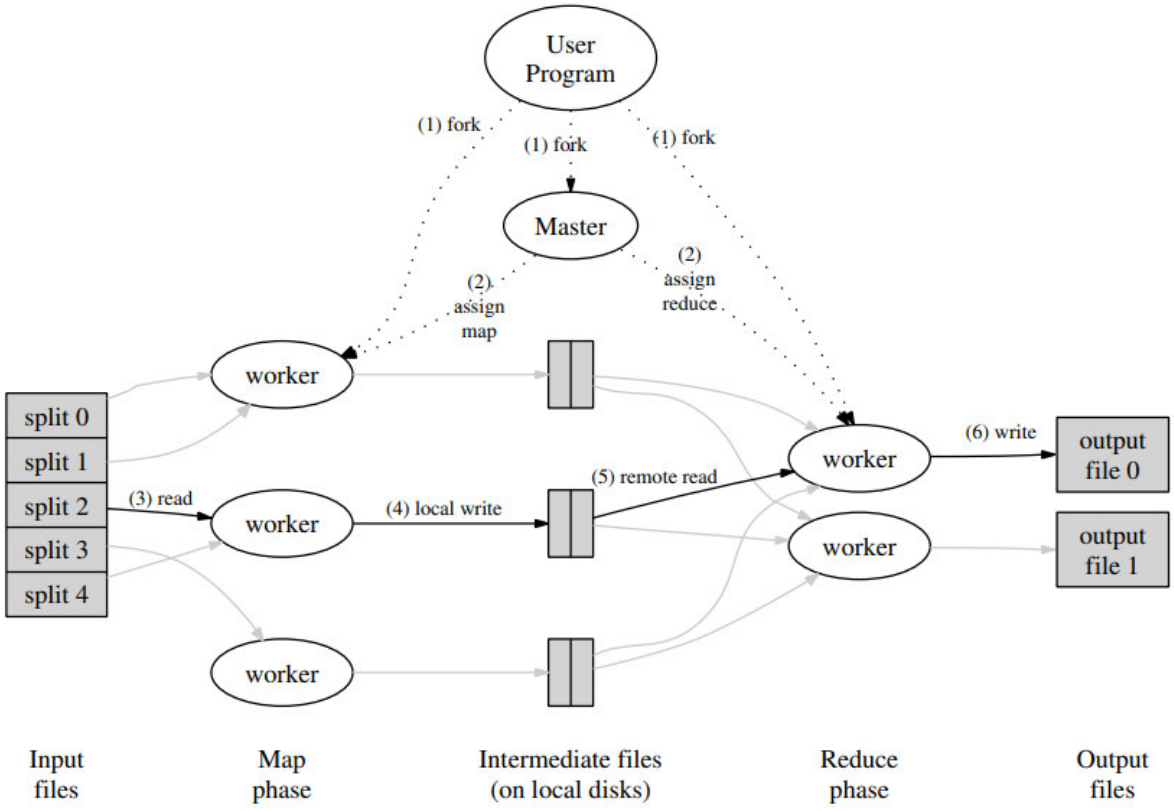
disponibilizado abertamente para que seja utilizado por qualquer usuário. Neste capítulo, serão apresentadas as ferramentas e os *frameworks open source* utilizados neste trabalho, como o modelo de programação MapReduce e os *frameworks* Apache Hadoop e Apache Ignite.

2.2.1 MapReduce

MapReduce é um modelo de programação criado pela Google em 2014, desenvolvido para simplificar o processamento de dados em paralelo sobre uma estrutura de grandes *clusters* distribuídos. Ele é baseado nas primitivas *map* e *reduce*, utilizadas na programação funcional (DEAN; GHEMAWAT, 2014).

O MapReduce é altamente escalável, tolerante a falhas e permite o balanceamento de carga entre os *clusters*. Os programas baseados em MapReduce possuem duas funções básicas, *map* e *reduce*. A função *map* recebe um par chave/valor e os combina de acordo com um critério estabelecido pelo usuário. A partir disso, ele gera um conjunto intermediário de dados do mesmo formato. Esse conjunto é passado para a função *reduce*, que agrupa todos os valores associados com a mesma chave intermediária (DEAN; GHEMAWAT, 2014). Na Figura 3 é possível visualizar uma visão geral da execução do MapReduce.

Figura 3 – Visão global de uma execução MapReduce



Fonte: (DEAN; GHEMAWAT, 2014)

Em uma visão geral da execução do MapReduce, primeiramente a biblioteca MapReduce divide o dado de entrada em várias partes. Em seguida, o biblioteca executa várias cópias do programa sobre o conjunto de máquinas do *cluster*, sendo uma delas eleita como "nó" mestre, que designa as tarefas *map* e *reduce* para os "nós" trabalhadores. Quando um "nó" trabalhador é designado para uma tarefa *map*, esse "nó" lê a entrada de dado correspondente e passa o par chave/valor para a função *map*, definida pelo usuário. O conjunto de pares intermediários chave/valor são armazenados na memória e escritos no disco em várias partes, a partir de uma função de particionamento definida pelo usuário. Por último, a localização dos conjuntos de pares chave/valor intermediários são devolvidos para o "nó" mestre, que encaminha as informações para os nós *reduce*.

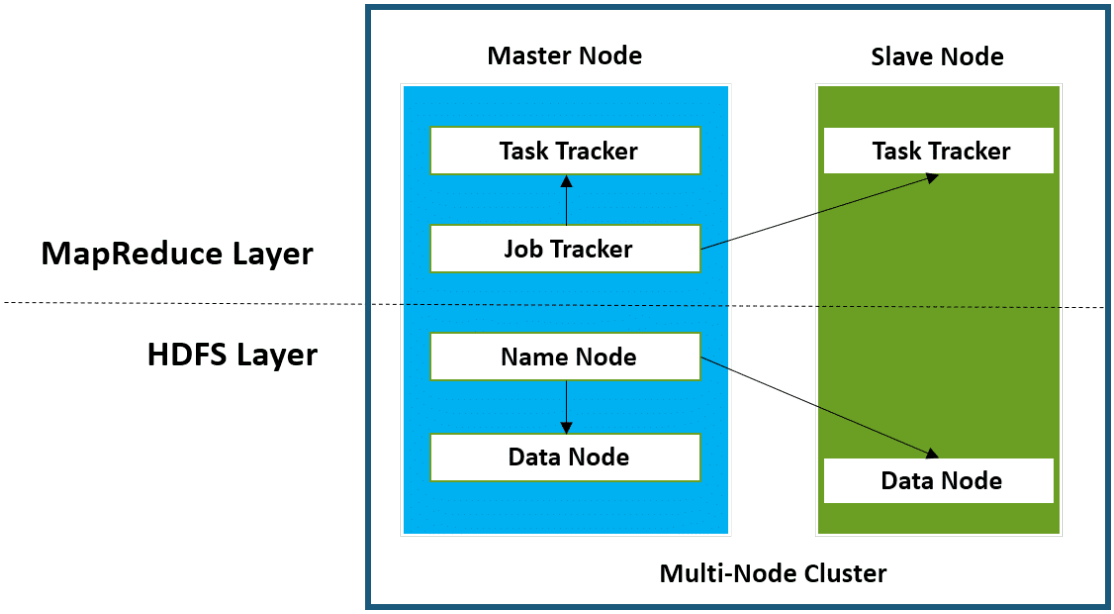
O modelo de programação MapReduce é utilizado em muitas aplicações pela Google. Entre essas aplicações estão: a mineração de dados, aprendizado de máquina, geração de dados para serviço de busca na *web*, entre outras (DEAN; GHEMAWAT, 2014). A implementação do modelo MapReduce utilizada pela Google é software proprietário, porém, seu modelo foi inspiração para muitos *frameworks open source*, assim como o Apache Hadoop e o Apache Ignite, que serão abordados nos próximos subtópicos.

2.2.2 Apache Hadoop

Como dito no tópico anterior, apesar do modelo MapReduce disponibilizado pela Google ser um *software* proprietário, há implementações *open source* disponíveis, uma dessas implementações é o Apache Hadoop (APACHE HADOOP, 2019).

O Apache Hadoop é uma implementação *open source* do MapReduce desenvolvida pela Apache, com uma arquitetura semelhante a da implementação feita pela Google. Como a implementação do Apache Hadoop é semelhante a do MapReduce, ambos possuem as mesmas características, a principal diferença está entre os seus sistemas de arquivos. No caso da Google, é utilizado o GFS (*Google File System*), já no caso do Hadoop, os dados encontram-se distribuídos no *Hadoop Distributed File System* (HDFS).

Figura 4 – Arquitetura interna do Apache Hadoop



Fonte: (APACHE HADOOP, 2019)

Na Figura 4 observamos a visão global da arquitetura interna do Hadoop, que é dividido em duas camadas. Na camada do sistema de arquivos, *Hadoop File System* (HDFS), o conjunto de dados de entradas são distribuídos em um conjunto de máquinas. Esse sistema de arquivos cria cópia de dados e redistribui em computadores próximos ao *cluster* para aumentar a confiabilidade. O HDFS possui dois processos principais, o *namenode* e *datanode*. O *namenode* é executado em uma única máquina mestre. Nele contêm informações sobre as máquinas no *cluster* e os detalhes sobre os blocos de dados persistidos nas máquinas componentes do *cluster*. O *datanode* processa a execução em todas as máquinas do *cluster*, comunicando com o *namenode* para saber quando buscar dados no disco rígido local.

Na camada MapReduce há um *JobTracker* e um número de processos *TaskTracker*. O *JobTracker* executa na mesma máquina que o *namenode*. Os usuários enviam seus *jobs* para o *JobTracker* que divide a tarefa entre as máquinas do *cluster*. Cada máquina no *cluster* executa um processo *TaskTracker*, que se comunica com o *JobTracker*, que designa uma tarefa *map* ou *reduce* quando possível.

O uso do Hadoop é indicado para aplicações que realizam processamento offline de grandes lotes de históricos e/ou *payloads* analíticos, em que a latência e transações não são importantes.

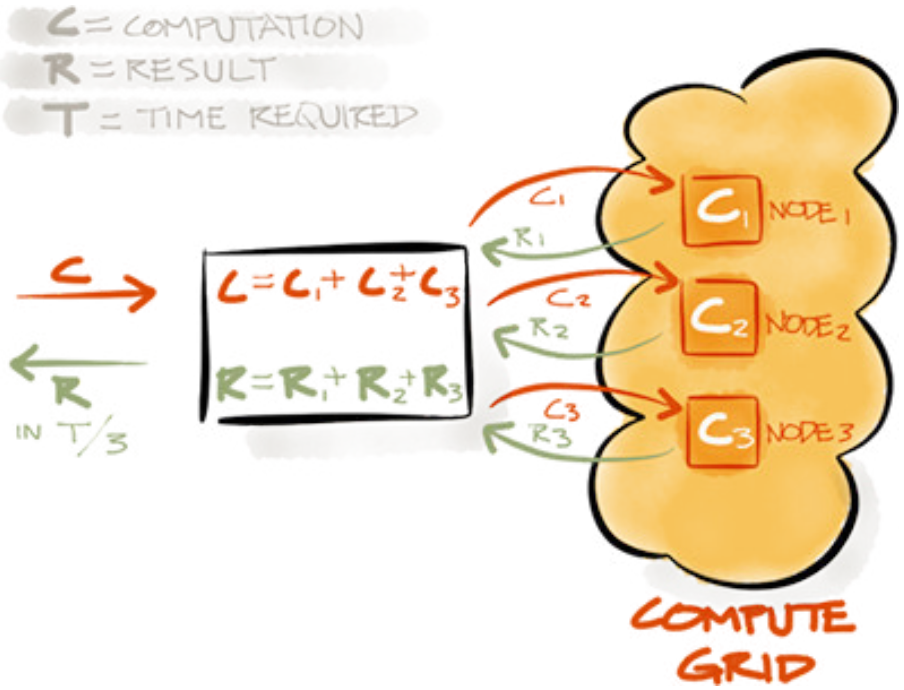
2.2.3 Apache Ignite

O Apache Ignite é um *framework* distribuído, integrado, de alta performance baseado em *Java Virtual Machine* (JVM), oferecendo rapidez e alta escalabilidade para o processamento de grande conjunto de dados (APACHE IGNITE, 2019). O Apache Ignite originou-se de uma doação da *GridGain Systems*, em 2014, para a *Apache Software Foundation* sob o projeto *open source* Apache Ignite.

O Apache Ignite utiliza o conceito *In-memory Computing*, que consiste no gerenciamento de um grande volume de dados com rapidez e economia de recursos. Em um paradigma convencional de armazenamento, os dados ficam salvos no disco rígido, que demanda muito tempo para o acesso aos dados, principalmente quando se trata de um grande conjunto de dados, no caso do *Big Data*. No *In-memory Computing*, os dados ficam salvos na RAM, tornando o tempo de acesso mais rápido. Em sua API, o Apache Ignite repassa o processamento para os *grids* de memória, resultando na rapidez em suas operações.

A API do Apache Ignite possui suporte à utilização de diferentes bancos de dados, como RDBMS, *NoSQL*, Hadoop Data Stores e diversas plataformas em nuvem. Além disso, possui suporte a várias linguagens de programação, como SQL, PHP, Java, Scala, entre outras. Na Figura 5 é possível visualizar o funcionamento do Apache Ignite MapReduce, conceito abordado no subtópico 2.2.1, em um esquema *In-Memory Compute Grid*, um dos conceitos do *In-Memory Computing*.

Figura 5 – Modelo de execução *In-Memory Compute Grid*

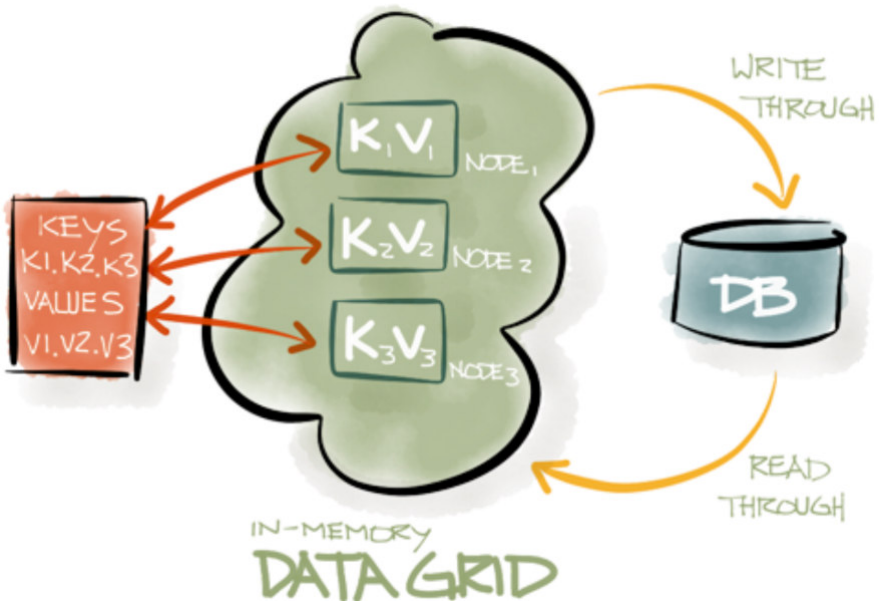


Fonte: (APACHE IGNITE, 2019)

O Apache Ignite *In-Memory Data Grid* (IMDG), outro conceito do *In-Memory Computing*, é um sistema de armazenamento de dados na memória RAM, similar a um *hash map*, onde os dados são armazenados em um sistema de chave/valor. A diferença entre o IMDG e outros sistemas de armazenamento é que ele é baseado em *caching* distribuído, podendo transacionar e atualizar os dados em tempo real, diferentemente do HDFS.

Na Figura 6 é ilustrado o funcionamento do IMDG, onde há um conjunto de chaves {k1, k2, k3} onde cada chave corresponde a um "nó" diferente. Caso haja um banco de dados externo, o IMDG irá se conectar para que possa ler e escrever no banco. (APACHE IGNITE, 2019).

Figura 6 – Modelo de execução *In-Memory Data Grid*



Fonte: (APACHE IGNITE, 2019)

O Apache Ignite foi escolhido para este trabalho pelo fato de exigir um hardware de baixo custo para um processamento de alta performance. Além disso, foi desenvolvido na linguagem de programação Java, disponibilizando uma vasta documentação e permitindo uma grande facilidade de manutenção, contando com exemplos de código fonte para diferentes tipos de uso.

2.3 Análise de Sentimento

Segundo Liu (2012), a análise de sentimento ou mineração de opinião é um campo de estudo que analisa a opinião das pessoas, assim como suas emoções, atitudes, comportamentos e avaliações. Essa análise pode estar ligada a produtos, serviços, organizações ou indivíduos.

Essa definição de Análise de Sentimento abrange muitas tarefas, análise de sentimento, mineração de opinião e mineração de *reviews* são algumas das análises que se enquadram na conjunto da análise de sentimento.

A análise de sentimento obteve um grande crescimento nos últimos anos devido a proliferação de aplicações comerciais que promoveram um grande aumento o interesse de pesquisas voltadas a essa área. Consequentemente a isso, outro fator importante para o crescimento da análise de sentimento é a grande quantidade de problemas que ainda devem ser pesquisados (LIU, 2012).

2.3.1 Definição de Opinião

Uma opinião é definida como uma quádrupla (g, s, h, t), onde g é a opinião alvo, s é o sentimento em relação ao alvo, h é o dono da opinião e t é o tempo em que a opinião é expressada. Essa definição não é utilizada na prática, pois quando se trata de *reviews* online de produtos e serviços, por exemplo, a descrição completa do alvo em questão pode não estar na mesma sentença.

Para que essa opinião possa ser definida de maneira completa, surge um quinto elemento, a entidade e. Essa entidade é um serviço, tópico ou pessoa descrito com um par e: (T,W), onde T é uma hierarquia de partes e subpartes e W é o conjunto de atributos de e, além disso, cada subparte também possui o próprio conjunto de atributos.

Tendo essa decomposição realizada, podemos definir uma opinião como uma quíntupla (e_i, a_{ij}, s_{ijkl}, h_k, t_l), onde e_i é o nome de uma entidade, a_{ij} é um aspecto de e_i, s_{ijkl} é o sentimento no aspecto a_{ij} da entidade ei, h_k é o detentor da opinião e t_l é o tempo em que a opinião é expressada por h_k (LIU, 2012).

A descrição dos elementos é dada por:

- **Entidade:** é uma pessoa, marca, instituição, serviço, etc. Uma entidade é um par e(T, W), onde T é a hierarquia de componentes e subcomponentes e W é o conjunto de atributos de e.
- **Aspecto:** aspectos são os componentes e atributos da entidade e.
- **Nome de aspecto e Expressão de aspecto:** o nome de aspecto é o nome que é associado ao aspecto, já a expressão de aspecto é a palavra real ou frase que indicou o aspecto no texto.
- **Nome de entidade e expressão de identidade:** o nome de entidade é o nome dado a entidade pelo usuário. A expressão de identidade é a palavra real ou frase que apareceu no texto indicando a entidade.

- **Dono da opinião:** é o locutor que expressou a opinião, podendo ser uma pessoa ou organização.

O sentimento s_{ijkl} pode ser positivo, negativo, neutro ou expressado por níveis de intensidade distintos. O nível de intensidade distinto pode ser observado em pesquisas de opinião e questionários, onde, partindo de uma determinada opinião ou afirmação, é possível assinalar diferentes tipos de resposta, como "concordo" ou "concordo plenamente".

2.3.2 Tipos de Opinião

Segundo Dubey e Gupta (2016), existem dois tipos de opiniões, opiniões regulares e opiniões comparativas. Opiniões regulares podem ser divididas em dois subtipos: opinião direta e opinião indireta. Nas opiniões regulares, a opinião é expressada diretamente a uma entidade. No caso da opinião regular direta, a entidade é bem definida, como por exemplo, na frase "Aquela calça é bonita". Já no caso da opinião regular indireta, há uma certa dificuldade em identificar a entidade principal, pois mais de uma entidade pode ser considerada.

Quando se trata de opiniões comparativas, há também dois tipos: opiniões explicitas e implícitas. Uma opinião explicita é uma afirmação subjetiva que dá uma opinião regular ou comparativa. Um exemplo de opinião explicita está na frase: "A TV da LG é melhor que a da Samsung". Já uma opinião implícita poderia ser caracterizada por uma afirmação objetiva que implica uma opinião regular ou comparativa.

2.3.3 Desafios da Análise de Sentimentos

Kolkur, Dantal e Mahe (2015) cita alguns desafios da análise de sentimentos em seu trabalho, como o sarcasmo, negação e comparações. Além disso, Silva (2013) acrescenta outro, o conhecimento de mundo. Adiante, serão citados os principais desafios da análise de sentimentos.

- **Sentimentos implícitos e sarcasmos:** nem todas as sentenças possuem sentimentos explícitos, é comum que sentenças possuam sentimentos expressados de forma implícita, como no exemplo: "O jogo do flamengo está muito parado. Não sei como estou conseguindo assistir". Apesar da sentença não possuir palavras claras expressando um sentimento negativo, é perceptível que ela possua tal sentimento. Como a entonação expressada na sentença não pode ser detectada, o jogo de palavras no sarcasmo pode induzir ao erro durante a análise de sentimentos.
- **Comparações:** o principal desafio no momento de uma comparação é analisar qual é a entidade principal na sentença. Por exemplo: "O carro da Fiat é melhor que o da Renault". Na sentença, há a palavra "melhor", o que configura uma opinião positiva

caso a entidade principal seja o carro da Fiat. Porém, caso a entidade principal seja o carro da Renault, essa sentença teria o valor negativo.

- **Negação:** a negação é o principal desafio da análise de sentimentos, pois nem sempre elas estão explicitas. Na sentença "Eu não gosto de quiabo", o sentimento de negação é claro. Já na sentença "Ela não só gostou do presente, como também amou as lembranças.", apesar de ter o operador de negação "não", ele não altera o sentimento da sentença, que é positiva.

- **Conhecimento de mundo:** toda linguagem tem suas peculiaridades e expressões regionais, principalmente no caso da língua portuguesa, onde o regionalismo dentro do Brasil é bastante presente. Na sentença: "Essa festa tá aziada" é relatado um sentimento negativo, mas, para identifica-lo é necessário que haja um conhecimento de gírias e regionalismos.

2.3.4 Níveis da Análise de Sentimentos

A análise de sentimentos possui três níveis, são eles: sentença, aspecto e documento (LIU, 2012) .

- **Análise a nível de aspecto e entidade:** é uma análise que não foca nas construções de linguagem, como sentenças, documentos e parágrafos. Esse tipo de análise é construída com base no alvo da opinião. Esse tipo de análise é bastante comum no contexto de revisões de produtos e serviços.

- **Análise a nível de documento:** Nessa análise, um documento inteiro é classificado com um sentimento positivo ou negativo. Esse tipo de análise assume que um documento expressa uma opinião sobre uma única entidade (PANG et al., 2002). Por exemplo: seria analisado se um documento é positivo ou negativo em relação a sua entidade, no caso de um documento que forneça opiniões sobre um filme.

- **Análise a nível de sentença:** Na análise em nível de sentença, uma determinada expressão é classificada como positiva, negativa ou neutra. Nesse nível, a análise está fortemente relacionada com a classificação da subjetividade. Nela, é possível fazer a distinção de sentenças que objetivas (que expressam fatos) de sentenças subjetivas (que expressam pontos de vista) (WIEBE, 1999). Esse tipo de análise será utilizada no trabalho.

2.3.5 Técnicas para Análise de Sentimentos em Nível de Sentença

Há três tipos de técnicas que podem ser utilizadas para a classificação de opiniões, são elas: o aprendizado de máquina, léxico e abordagem híbrida. Para Pang e Lee (2008),

léxico e aprendizado de máquina são os dois métodos principais. Os métodos baseados em léxico utilizam dicionários, como por exemplo o *SentiWordNet*. Em técnicas baseadas em aprendizado de máquina, costumam-se utilizar o *Support Vector Machine* (SVM) e o *Naïve Bayes* como classificadores. Já a abordagem híbrida mescla a abordagem léxica com o aprendizado de máquina. A seguir, serão detalhadas as duas principais técnicas para Análise de Sentimentos:

- **Técnicas baseadas em léxico:** as técnicas baseadas em léxico realizam uma classificação baseada em dois sentimentos: positivo e negativo. Nessa técnica, um dicionário de palavras é utilizado para que o valor de determinada palavra possa ser atribuído como positivo ou negativo. Primeiramente, o valor do sentimento de cada palavra é definido, em seguida, há uma função de combinação para que possa ser feita a previsão final da frase com base na classificação das palavras que a compõe. De acordo com Pang e Lee (2008), há três maneiras de construir um dicionário léxico, são elas: a construção manual, métodos baseado em dicionário e métodos baseado em corpus. Na construção manual, o usuário precisa definir, manualmente, o peso ou sentimento referente a cada palavra do dicionário. No método baseado em dicionário, as palavras são buscadas em um dicionário e, desse modo, a pontuação de cada palavra é atribuída de acordo com a mesma presente no dicionário. Por fim, nas técnicas baseada em corpus, já há um dicionário pré-definido com valores positivos e negativos para as palavras, porém, são usadas contagens de palavras e outras medidas de incidência e frequência de palavras para que possam ser classificadas as opiniões. Esse último método permite uma alta acurácia.

- **Técnicas baseadas em aprendizado de máquina:** As técnicas baseadas em aprendizado de máquina pertencem às técnicas de classificação supervisionadas. Nesse tipo de técnica são utilizados dois tipos de documentos, os documentos de treino e os documentos de teste (PANG; LEE, 2008). O conjunto de treino é usado por um classificador automático para aprender a diferenciação de característica de documento e o conjunto de teste é utilizado para verificar o quão bem o classificador executa. Os classificadores mais utilizados são: *Naïve Bayes*, *Maximum Entropy* e *Suport Vector Machine* (SVM). Na etapa de aprendizado de máquina, um conjunto de dados coletados são colocados em treino para que, em seguida, os dados são submetidos ao classificador para treino. O conjunto de treino é utilizado para que o classificador aprenda a diferenciar o conjunto de características do documento e o de teste para testar o classificador em um conjunto maior de dados.

Neste capítulo, foram discorridas as principais definições sobre *Big Data*, Apache Ignite e análise de sentimento, essenciais para o entendimento do trabalho.

No capítulo a seguir, os conceitos serão aplicados no desenvolvimento da metodologia da análise de sentimentos do trabalho, por meio de um estudo de caso.

3 Estudo de Caso

Neste capítulo, será apresentada a metodologia proposta para este trabalho através de um estudo de caso em que serão aplicados os conceitos vistos nos capítulos anteriores.

Os softwares e hardwares utilizados no trabalho serão apresentados e detalhados no tópico 3.1. Já a metodologia proposta no trabalho será apresentada no tópico 3.2, onde serão conhecidas as etapas a serem seguidas durante a execução do trabalho. A primeira etapa a ser abordada é a de Coleta de Dados, onde serão coletados *tweets* referentes às queimadas na região amazônica que ocorreram em setembro de 2019. Em seguida, na etapa de Pré-Processamento, serão conhecidos os ajustes realizados na base de dados. Posteriormente, na etapa de Análise de Sentimentos, a metologia usada no trabalho será apresentada. Por último, na etapa de Discussão dos Resultados, os resultados da aplicação serão discutidos e comparados entre si, obtendo-se uma análise detalhada do trabalho executado.

3.1 Software e Hardware utilizados

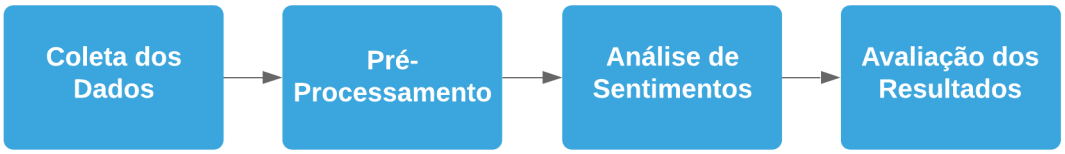
Para a captura dos *tweets* sobre o tema, foi utilizada a linguagem de programação Python, através da biblioteca *Tweepy* (TWEEPY, 2019). Os dados extraídos do Twitter foram salvos em um arquivo de texto onde, em cada linha, um *tweet* era armazenado.

Para a análise de sentimento, a ferramenta criada foi desenvolvida na linguagem de programação Java, utilizando a IDE Eclipse Neon, disponível no site *Eclipse Foundation* (ECLIPSE FOUNDATION, 2019). A *framework* Apache Ignite foi incorporada ao código através da API Java obtida em seu site oficial (APACHE IGNITE, 2019). A versão da IDE utilizada foi a Neon.3 *Release* (4.6.3) e a do *Framework* Apache Ignite foi a 2.7.5. O Hardware utilizado no desenvolvimento e testes foi um *notbook* DELL, processador Intel Core I5, 4GB de memória RAM, 1TB de HD e sistema operacional Windows 10.

3.2 Metodologia Proposta

A metodologia proposta no desenvolvimento do trabalho é ilustrada na Figura 7. As etapas que foram seguidas no desenvolvimento foram: a coleta dos dados; o pré-processamento; a análise de sentimentos; e, a avaliação dos resultados.

Figura 7 – Etapas da metodologia proposta



Fonte: Elaborada pelo autor

3.2.1 Coleta dos Dados

Esta primeira etapa corresponde à etapa onde será formada a base de dados utilizada na aplicação. A formação dessa base de dados foi composta através da extração de *tweets*. O Twitter é uma rede social onde os usuários podem expressar as suas opiniões através de mensagens de até 280 caracteres. Devido ao número de caracteres limitado em comparação com outras redes sociais, a escolha do Twitter se mostra muito eficaz, pois com um número limitado de caracteres, o usuário busca expressar suas opiniões de maneira objetiva.

O Twitter disponibiliza ferramentas para desenvolvedores através do *Twitter for Developers*, um site feito para desenvolvedores que desejam realizar pesquisas e fazer trabalhos através do Twitter (TWITTER DEVELOPER, 2019). No site pode ser encontrada uma vasta documentação sobre como ter acesso aos *tweets*, registro do desenvolvedor e APIs para a obtenção de *tweets*.

Dentre as APIs disponíveis para a obtenção de *tweets*, a opção escolhida para este trabalho foi a Twitter *search* API, pois demostrou ser fácil de manipular, criando-se uma aplicação em poucas linhas de código. A Twitter *search* API recupera *tweets* a partir de requisições HTTP, usando o método GET através do endereço http://search.twitter.com/search.json?q="parâmetro de busca". A Twitter *search* API retorna somente mensagens recentes, postadas em um período máximo de 7 dias anteriores ao dia da busca, passando-se um parâmetro a cada busca.

Os *tweets* utilizados para compor a base de dados foram capturados entre os dias 2 e 10 de setembro de 2019, referentes a esse mesmo período de tempo. Como a API Twitter *Search* só aceita um parâmetro por vez em suas buscas, foram realizadas várias requisições. Em algumas requisições foram buscados *tweets* que continham a palavra "amazônia", em outras, foram buscados *tweets* que continuam a palavra "queimadas".

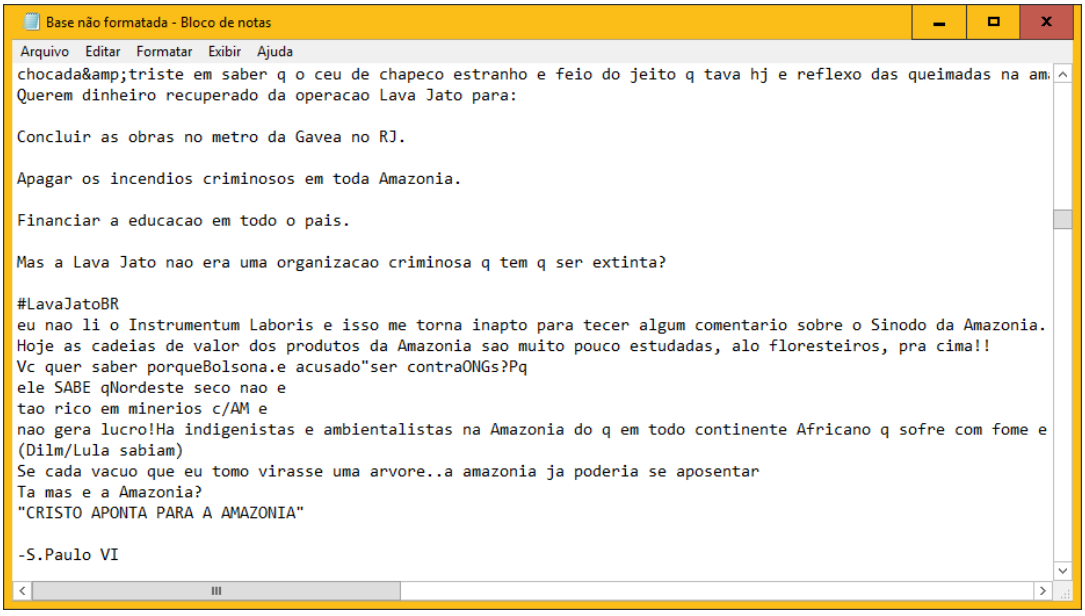
Durante as últimas semanas de agosto e início do mês de setembro, discussões sobre as queimadas na amazônia estavam bastante pertinentes no Twitter. Devido ao grande volume de *tweets* sobre o tema, a busca por *tweets* com as palavras "amazônia" e "queimadas" ocorreu sem dificuldades.

A Twitter *Search* API retorna um arquivo JSON contendo várias informações sobre o *tweet*, como a data de criação, id e outras informações sobre o *tweet*, porém apenas o conteúdo dos *tweets* é realmente necessário para uso neste trabalho. O conteúdo dos *tweets* foram salvos em um arquivo de texto. Conforme os *tweets* foram sendo extraídos através da aplicação com o uso do *Tweepy*, cada *tweet* foi sendo gravado em uma linha em um arquivo .txt e, logo após o fim da requisição, o arquivo era fechado. Ao todo, foi contabilizado uma quantidade total de 2564 *tweets* obtidos.

3.2.2 Pré-Processamento

Apesar de que, no momento da captura dos *tweets*, apenas o conteúdo do *tweet* tenha sido coletado, sem a presença de links de imagens, menções ou de *retweets*[1], o pré-processamento se mostra essencial como etapa, pois, sem ela, não só *tweets* com falhas na formatação poderiam acarretar em erros durante o processamento, como também uma classificação incorreta de determinados *tweets* poderia ser realizada. Mesmo com os arquivos apresentando apenas o conteúdo dos *tweets* dispostos em cada linha do arquivo de texto, alguns *tweets* apresentaram quebra de linha, tornando necessário um pré-processamento manual, organizando os *tweets* para que cada um pudesse ser disposto em uma só linha do arquivo, o que facilitará durante a etapa de análise de sentimentos. Além da retirada de quebras de linha, alguns *tweets* repetidos ou que fugiam muito do tema proposto foram retirados. Na Figura 8 é possível observar um trecho da base de dados ainda não formatada, com várias quebras de linhas entre os *tweets*.

Figura 8 – Base não formatada



Fonte: Elaborada pelo autor

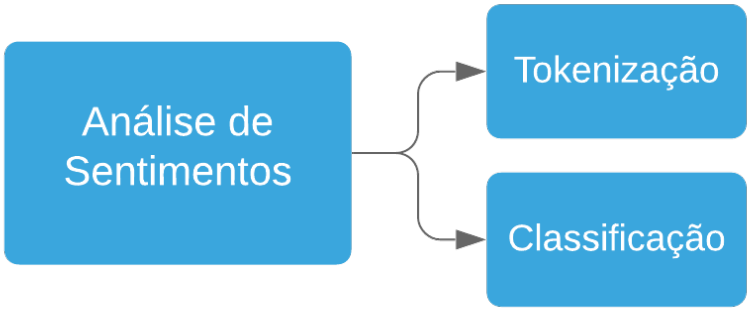
[1] Republicações de um *tweet*

Com algumas exceções, o conteúdo da maioria dos *tweets* não foi alterado, assim como a quantidade dos mesmos. Logo, após a fase de pré-processamento, obteve-se um total de 2481 *tweets*, que foram classificados manualmente como positivos, negativos ou neutros, sendo atribuída a expressão "pos", "neg" e "neu", respectivamente, no início de cada *tweet*. Essa classificação será importante para análise futura dos resultados, pois assim, será possível comparar com outras avaliações que serão apresentadas nas seções seguintes.

3.2.3 Análise de Sentimentos

Na etapa de análise de sentimentos ocorrerá, de fato, a análise sobre os *tweets* coletados e pré-processados. A análise de sentimentos ocorrerá em nível de sentença e, para que ocorra de tal modo, ela será realizada fazendo-se o uso do modelo de programação *MapReduce*, através do *Framework* Apache Ignite. Nesse modelo de programação, a análise de sentimentos ocorrerá em duas etapas: a tokenização, correspondente ao *map*, onde os *tweets* terão cada palavra classificada com seu respectivo valor e *reduce*, onde o somatório das palavras de cada *tweet* definirá o valor final e sentimento da frase. Dessa forma, com uma tarefa para a atribuição de valores às palavras e outra para a classificação dos *tweets*, ao todo, de maneira paralelizada, esse modelo de programação se mostra eficaz quando se trata de uma análise em nível de sentença. Na Figura 9 é possível observar as subdivisões da etapa de Análise de Sentimentos, que serão abordadas adiante.

Figura 9 – Subdivisões da Análise de Sentimentos



Fonte: Elaborada pelo autor

3.2.3.1 Tokenização

Nesta etapa, a classificação de cada palavra do *tweet* será realizada através da atribuição de um valor positivo, negativo ou neutro para cada palavra. Esses valores serão atribuídos de acordo com os valores pertencentes às palavras em cada dicionário léxico

carregado na aplicação. Caso a palavra analisada tenha o seu valor negativo igual ao valor positivo, essa palavra será interpretada como de valor neutro, não relevante. Nessa aplicação, três dicionários léxicos diferentes serão utilizados. As particularidades de cada um serão explicadas adiante.

Para uso na aplicação, os três dicionários léxicos escolhidos foram: *OpLexicon* (OPLEXICON, 2019), *SentiLex* (SENTILEX, 2019) e *SentiWordNet* (SENTIWORDNET-PT-BR, 2019). Em português, há poucos dicionários léxicos disponíveis. A escolha dos três dicionários foi feita com base na popularidade dos mesmos. O dicionário léxico *SentiWordNet PT-BR* é baseado no dicionário léxico de mesmo nome (*SentiWordNet*), que é baseado no dicionário *WordNet*, onde as palavras em inglês são agrupadas em conjuntos de sinônimos (*synsets*) mantendo o mesmo padrão de pontuação do dicionário *SentiWordNet*. A versão do *SentiWordNet PT-BR* utilizada neste trabalho é a 1.0. O padrão de pontuação consiste em um identificador baseado na identificação internacional baseada no *WordNet 3.0*, uma pontuação positiva e negativa da palavra, ambas variando entre 0 e 1, e a palavra em si. Na Figura 10 é mostrado um exemplo de registro do dicionário *SentiWordNet*.

Figura 10 – Registro do dicionário *SentiWordNet*

```
00001740-a      0.125   0       capaz
```

Fonte: (SENTIWORDNET-PT-BR, 2019)

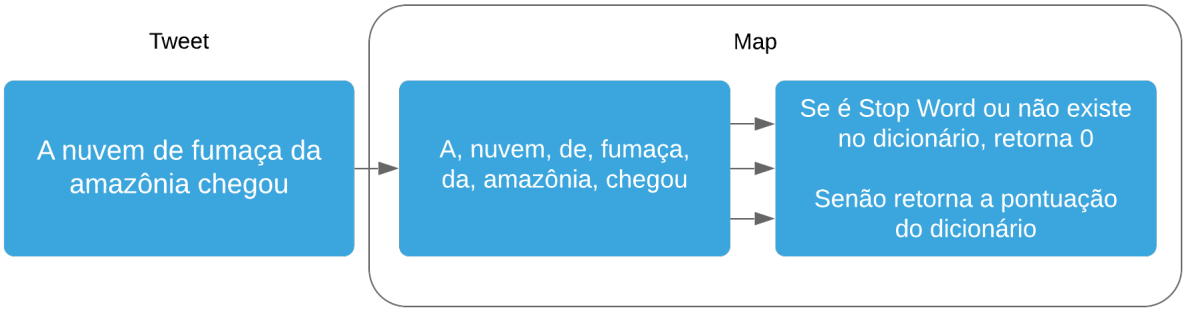
Os dicionários *SentiLex* e *OpLexicon* mantêm o mesmo padrão de pontuação um do outro, onde cada palavra possui somente um valor, sendo 0 para palavra neutra, 1 para palavra positiva e -1 para palavra negativa. Adiante será explicado o método para determinar o valor de cada *tweet* de acordo com cada dicionário.

Antes de iniciar a aplicação, os "nós" trabalhadores da *grid* do *Framework* Apache Ignite deverão ser inicializados em conjunto. Para que possam ser inicializados, deve-se executar o comando "ignite.bat", presente no diretório raiz do *Apache Ignite*. Logo após a inicialização dos "nós" trabalhadores, a aplicação deverá ser inicializada. Quando a aplicação for inicializada, os três dicionários léxicos são carregados na memória, assim como o arquivo contendo a base dos *tweets* já pré-processados. A aplicação lê um *tweet* por vez e passa cada um para o método *map*, onde ocorrerá a tokenização e a associação do *tweet* à sua pontuação.

Após receber o *tweet*, o "nó" mestre é responsável por separar as palavras e criar um *job* filho para cada uma das palavras. Logo após esse processo, as palavras são repassadas para os "nós" trabalhadores. Esses "nós" são responsáveis por atribuir os valores à palavra de acordo com o dicionário em uso. Por fim, cada "nó" retorna uma estrutura *map* contendo o resultado do processamento do *job*, que é enviado para outro "nó" que possa realizar a tarefa *reduce*. Caso não haja a ocorrência da palavra no dicionário em questão, essa

palavra não será classificada. Assim que todos as palavras do *tweet* forem processadas, inicia-se a tarefa *reduce*. Na Figura 11 a tarefa *map* é ilustrada de maneira simplificada, onde o *tweet* é dividido entre todas as suas palavras e a pontuação de cada palavra é atribuída de acordo com o dicionário em questão. Caso a palavra não exista no dicionário ou seja *Stop Word* (palavra sem relevância para os resultados, como por exemplo: a, o, de, para), a pontuação atribuída a ela é 0, caso tenha valor, esse valor é atribuído à palavra.

Figura 11 – Tarefa *map*



Fonte: Elaborada pelo autor

3.2.3.2 Classificação

Na etapa *reduce*, a soma das pontuações das palavras é realizada pelo "nó" mestre, obtendo-se o valor total do *tweet*. O método *reduce* recebe como parâmetro o agrupamento dos valores de cada palavra para que possa ser atribuído o valor total do *tweet*.

Em ambos os dicionários utilizados neste trabalho o intervalo de classificação do *tweet* poderá variar entre -1 e 1. Quanto mais próximo de 1, mais positivo é o *tweet*, quanto mais próximo de -1, mais negativo. Se o valor total é igual a 0, então esse *tweet* será classificado como neutro.

O cálculo da polaridade do *tweet* será realizado de forma semelhante entre os três dicionários, mas a atribuição dos valores de cada palavra será distinta entre o dicionário *SentiWordNet* e os dicionários *OpLexicon* e *SentiLex*. A metodologia do cálculo de polaridade é feita com base em uma média aritmética entre os valores totais das palavras (que é dado através da subtração do valor positivo com o valor negativo da palavra) pela quantidade de palavras relevantes, ou seja, que tenha um valor total diferente de 0. Na Figura 12 , analisaremos um *tweet* classificado como negativo com o uso do dicionário *SentiWordNet*.

Figura 12 – *Tweet* classificado como negativo com uso do *SentiWordNet*

```
>>> Tweet: 'neg Nao aguento mais essas queimadas
Palavra Processada: mais
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: não
Valor Positivo: 0.0
Valor Negativo: 0.625
Valor do tweet: -0.625 - Negativo
```

Fonte: Elaborada pelo autor

Na sentença "Nao aguento mais essas queimadas", a palavra "mais" tem o valor positivo e o valor negativo igual a 0, ou seja, não é uma palavra relevante e não entrará no cálculo de polaridade do *tweet*. Já a palavra "não" possui o valor positivo igual a 0 e o valor negativo igual a 0.625, tendo o seu valor total igual a -0.625 (subtração do valor positivo com o valor negativo). Como o *tweet* apresenta apenas uma palavra relevante, o valor total do *tweet* será apenas o valor da única palavra relevante, ou seja, o *tweet* possui valor -0.625, negativo.

Enquanto no *SentiWordNet* a pontuação final do *tweet* pode variar bastante entre -1 e 1, no *OpLexicon* e *SentiWord* a classificação final do *tweet* será -1, 0, ou 1, devido ao fato de que, nesses dicionários, cada palavra só terá um valor atribuído a ela, -1 para negativo, 0 para neutro e 1 para positivo.

O cálculo de polaridade de um *tweet* fazendo o uso dos dicionários *OpLexicon* e *SentiLex* ocorrerá do mesmo jeito que ocorre com o *SentiWordNet*. É realizada uma média aritmética entre os valores das palavras relevantes sobre a quantidade de palavras relevantes. Na Figura 13 observamos a classificação da polaridade de um *tweet* com o uso do *OpLexicon*.

Figura 13 – *Tweet* classificado com uso do *OpLexicon*

```
>>> Tweet: 'neg Enquanto isso estao encobrindo
o caso das pessoas que colocaram fogo na Amazonia em
Palavra Processada: encobrindo
Valor Positivo: 0.0
Valor Negativo: -1.0
Palavra Processada: fogo
Valor Positivo: 0.0
Valor Negativo: -1.0
Valor do tweet: -2.0 - Negativo
```

Fonte: Elaborada pelo autor

Como podemos observar na Figura 13, temos duas palavras relevantes, a palavra

"encobrindo" e a palavra "fogo". Realizando a soma das pontuações das palavras relevantes e dividindo pelo total de palavras relevantes, temos a pontuação total do *tweet*.

Para que houvesse uma maior acurácia na classificação dos *tweets* e, para que a quantidade de *tweets* não processados fosse reduzida, após a classificação dos tweets com os três dicionários léxicos foram propostos três conjuntos de melhorias nos dicionários léxicos em uso na aplicação. A primeira melhoria realizada nos três dicionários léxicos foi a adição de novas palavras. Foram adicionadas nos dicionários um conjunto de palavras que se mostravam presentes em uma grande quantidade de *tweets* mas não eram classificadas, pois não tinham um valor definido nos dicionários, assim como gírias, abreviações e palavrões. A segunda melhoria realizada foi a adição de verbos conjugados no gerúndio. Em uma grande quantidade, não só de *tweets* mas também de qualquer tipo de frase, é comum a presença de verbos conjugados no gerúndio, porém, nos três dicionários, apenas eram presentes os verbos no infinitivo. Assim, um *tweet* que continha um verbo conjugado no gerúndio não era classificado. Para todos os verbos, foram adicionadas as suas formas conjugadas no gerúndio contendo o mesmo valor do verbo correspondente presente no dicionário. A terceira melhoria foi a adição de palavras no tempo pretérito perfeito. Esse processo de adição foi feito de maneira semelhante a adição de verbos conjugados no gerúndio, porém foi feita apenas no dicionário *OpLexicon*, o único que não tinha verbos nesse tempo verbal.

Com exceção do primeiro conjunto de melhorias, em que foram adicionadas, manualmente, novas palavras nos três arquivos de texto referentes aos dicionários léxicos, a criação de palavras no gerúndio e de palavras no pretérito perfeito foi realizada através de uma funcionalidade implementada na aplicação. Nessa funcionalidade, eram lidos todos os verbos do dicionário em questão e, a partir disso, utilizou-se um método que foi implementado, responsável por criar o gerúndio e o verbo no pretérito do verbo lido. Ao final, esses novos verbos eram adicionados na lista que continha o restante de palavras do dicionário.

As melhorias foram adicionadas visando suprir deficiências dos dicionários, pois, antes das modificações, havia sido observado em vários *tweets*, certos conjuntos de palavras que não eram processadas pelos dicionários. Esses conjuntos eram, justamente, gírias, verbos conjugados no gerúndio e verbos conjugados no pretérito perfeito.

Na etapa seguinte, serão discutidos os resultados obtidos durante a fase de análise de sentimentos. Os resultados gerados a partir de todos os tipos de classificação realizados serão comparados e analisados detalhadamente na etapa de Avaliação dos Resultados.

3.2.4 Avaliação dos Resultados

Nesta etapa, os resultados obtidos na aplicação serão discutidos e comparados entre si. Para que a discussão dos resultados possa se tornar mais clara, exemplos individuais de *tweets* serão observados de acordo com algumas situações comuns durante o processamento. Serão analisados os resultados obtidos de acordo com a avaliação manual da base de dados, onde o sentimento de cada *tweet* foi atribuído de maneira manual pelo autor. Também serão analisados os resultados obtidos através da avaliação realizada pelos dicionários antes e depois das modificações. Essas avaliações serão feitas dessa forma pois, o sentimento verdadeiro dos *tweets*, determinado na avaliação manual, poderá ser comparado com os resultados das avaliações feitas através dos dicionários, assim, comparando o desempenho em relação aos acertos de cada um.

Durante a análise manual dos *tweets*, dentre os 2481 *tweets* analisados, 373 (15,04%) *tweets* foram classificados como positivos, 1586 (63,92%) como negativos e 522 (21,04%) como neutros, como mostrado na Tabela 1.

Tabela 1 – Avaliação Manual da Base de Dados

Sentimento	Classificação	Quantidade de Tweets
Positivo	15.04%	373
Negativo	63.92%	1586
Neutro	21.04%	522

Fonte: Elaborada pelo autor

O resultado do processamento de acordo com o dicionário *OpLexicon* foi de 603 (27,07%) *tweets* classificados como positivos, 1133 (50,87%) como negativos, 491 (22,4%) como neutros e 254 não processados, ou seja, que não houve qualquer palavra do *tweet* presente no dicionário para que fosse atribuída uma pontuação. No dicionário *SentiLex*, 456 *tweets* (24,38%) foram classificados como positivos, 1004 (53,69%) como negativos, 410 (21,04%) como neutros e 611 não foram processados. Já no dicionário *SentiWordNet*, 700 (28%) *tweets* foram classificados como positivos, 1377 (55,7%) como negativos, 395 (15,97%) como neutros e 9 não foram processados. Durante a análise dos resultados obtidos através da avaliação dos dicionários, os *tweets* não processados não serão avaliados, pois não há como atribuir um sentimento a um *tweet* que não recebeu nenhum tipo de pontuação. Nas Tabelas 2, 3, 4, estão representados, respectivamente, os resultados sobre os dicionários *OpLexicon*, *SentiLex* e *SentiWordNet* antes de quaisquer modificações.

Tabela 2 – Resultados *OpLexicon*

Sentimento	Classificação	Quantidade de Tweets
Positivo	27,07%	603
Negativo	50,87%	1133
Neutro	22,04%	491

Fonte: Elaborada pelo autor

Tabela 3 – Resultados *SentiLex*

Sentimento	Classificação	Quantidade de Tweets
Positivo	24,38%	456
Negativo	53,69%	1004
Neutro	21,04%	410

Fonte: Elaborada pelo autor

Tabela 4 – Resultados *SentiWordNet*

Sentimento	Classificação	Quantidade de Tweets
Positivo	28%	700
Negativo	55,70%	1377
Neutro	15,97%	395

Fonte: Elaborada pelo autor

Após as melhorias realizadas nos dicionários, o resultado do processamento de acordo com o dicionário *OpLexicon* foi de 602 (25%) *tweets* classificados como positivos, 1289 (52,84%) como negativos, 548 (22,46%) como neutros e 42 não processados, ou seja, que não houve qualquer palavra do *tweet* presente no dicionário para que fosse atribuída uma pontuação. No dicionário *SentiLex*, 554 *tweets* (23%) foram classificados como positivos, 1350 (56,06%) como negativos, 504 (20,93%) como neutros e 73 não foram processados. Já no dicionário *SentiWordNet*, 714 (29%) *tweets* foram classificados como positivos, 1369 (55,38%) como negativos, 389 (15,73%) como neutros e 9 não foram processados.

Nas Tabelas 5, 6, 7, estão representados, respectivamente, os resultados sobre os dicionários *OpLexicon*, *SentiLex* e *SentiWordNet* após as modificações.

Tabela 5 – *OpLexicon* Após Modificações

Sentimento	Classificação	Quantidade de Tweets
Positivo	25%	602
Negativo	52,84%	1289
Neutro	22,46%	548

Fonte: Elaborada pelo autor

Tabela 6 – *SentiLex* Após Modificações

Sentimento	Classificação	Quantidade de Tweets
Positivo	23%	554
Negativo	56,06%	1350
Neutro	20,93%	504

Fonte: Elaborada pelo autor

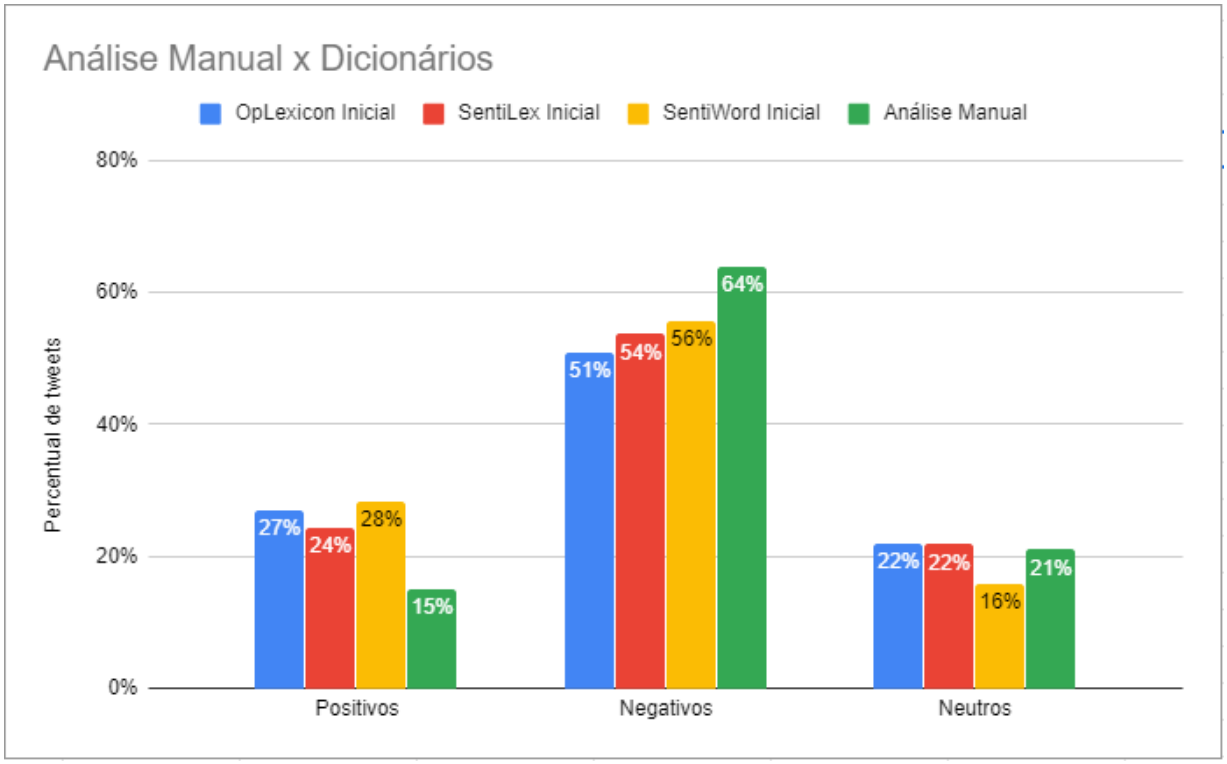
Tabela 7 – *SentiWordNet* Após Modificações

Sentimento	Classificação	Quantidade de Tweets
Positivo	29%	714
Negativo	55,38%	1369
Neutro	15,73%	389

Fonte: Elaborada pelo autor

Na Figura 14, observamos a comparação entre a análise manual dos resultados e a análise feita através dos três dicionários léxicos antes das modificações.

Figura 14 – Análise manual x Dicionários



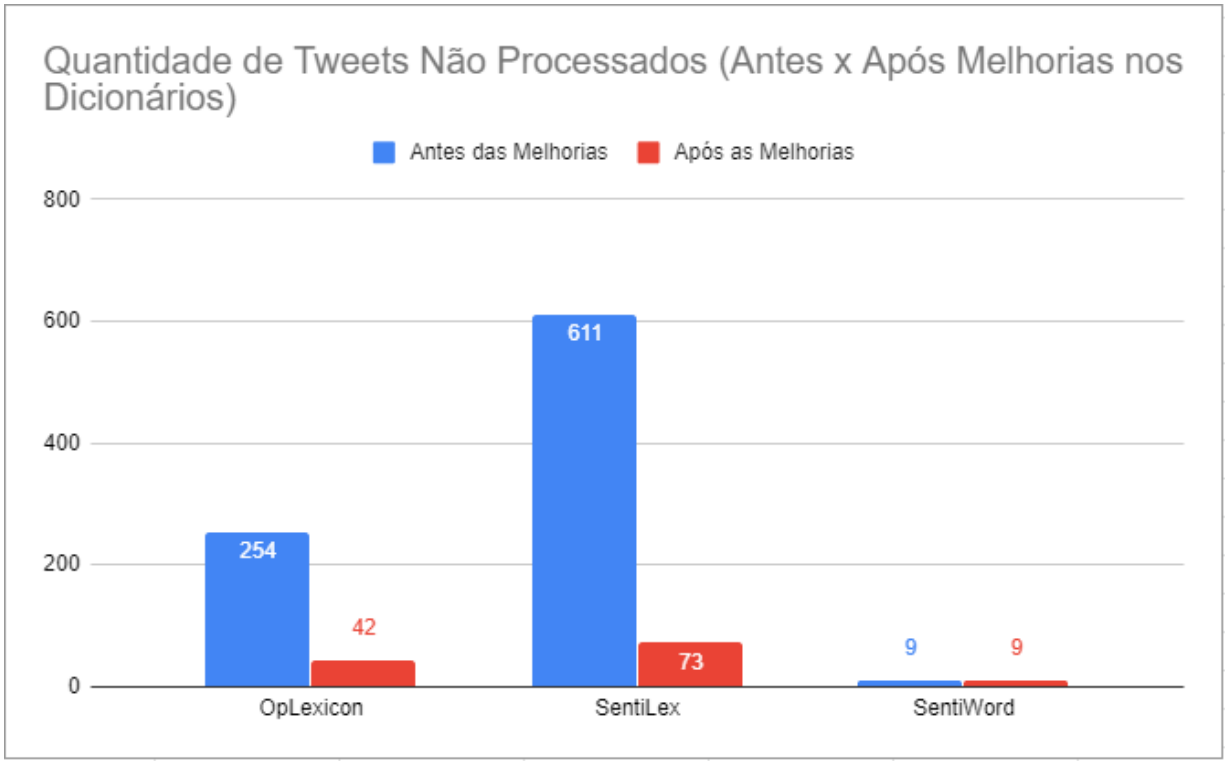
Fonte: Elaborada pelo autor

Analisando a Figura 14, observamos que, apesar da quantidade de *tweets* classificados como neutros na maioria dos dicionários ser próxima a quantidade de neutros classificados manualmente, a quantidade de *tweets* negativos não se manteve tão próxima, assim como a quantidade de positivos, que se manteve distante. Além disso, a comparação da Figura 14 exclui todos os *tweets* não processados no momento da análise feita com o uso dos dicionários.

Antes da melhoria dos dicionários, uma possível comparação entre os mesmos não seria adequada, pois a quantidade de *tweets* que não foram processados nos três dicionários eram bem distintas entre sim. Por exemplo, antes das modificações, enquanto o dicionario *SentiLex* tinha um total de 611 *tweets* não processados, o dicionário *SentiWordNet* tinha apenas 9 *tweets* não processados. Uma comparação apenas entre os *tweets* processados dos dois dicionários excluiria uma quantidade muito grande de *tweets* do dicionário *SentiLex*, o que poderia deslegitimar a comparação.

Após a melhoria dos dicionários, os três tiveram uma grande diminuição de *tweets* não processados. A quantidade de não processados entre os três dicionários passou a ser bem próxima. O dicionário *OpLexicon* apresentou 42 *tweets* não processados, o dicionário *SentiLex*, 73 e, o dicionário *SentiWordNet*, manteve a quantidade de *tweets* não processados, 9 *tweets*. Na Figura 15 podemos visualizar a diferença entre a quantidade de *tweets* não processados antes e depois nas modificações nos dicionários.

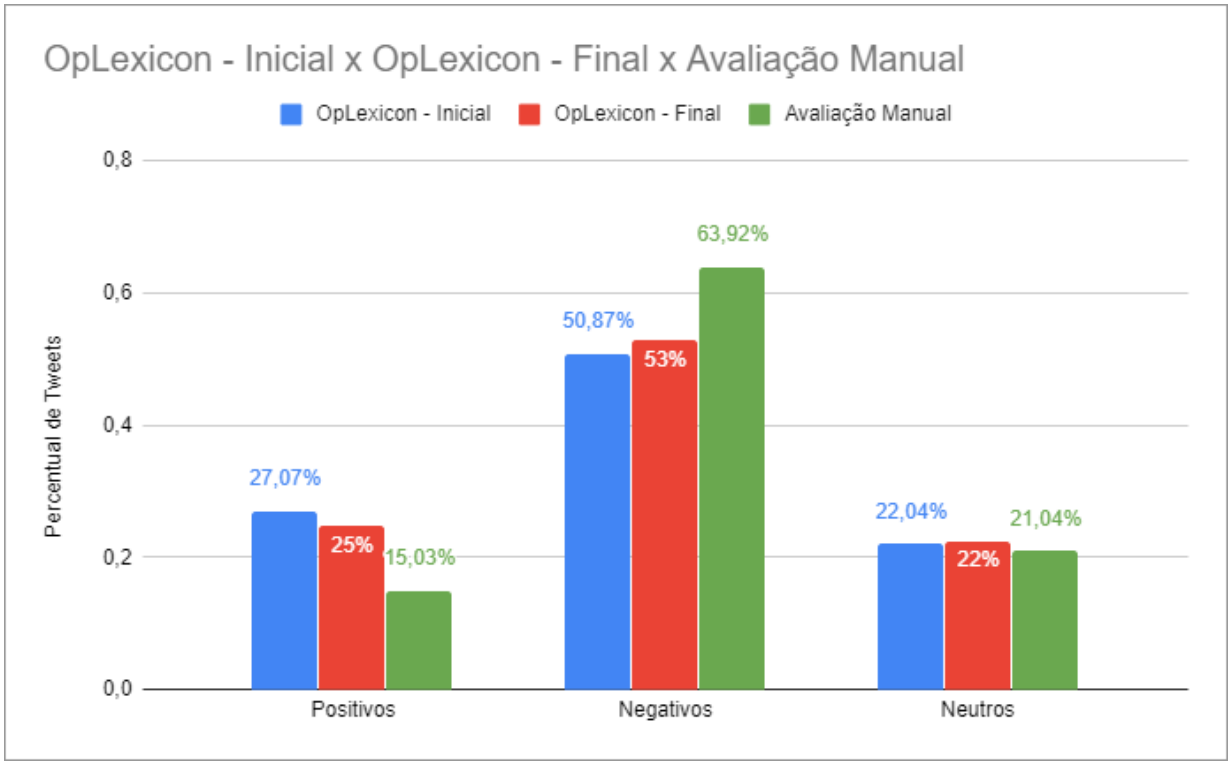
Figura 15 – *Tweets* não processados antes e depois das modificações nos dicionários



Fonte: Elaborada pelo autor

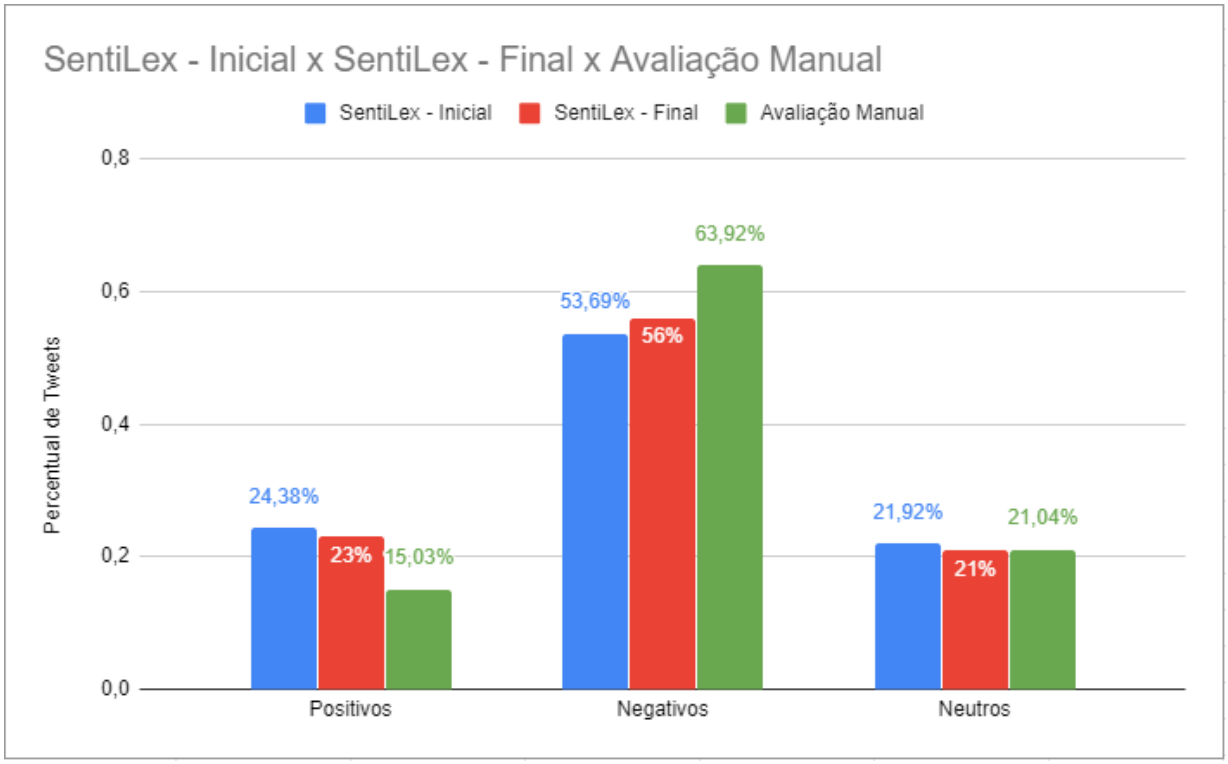
Nas Figuras 16, 17, 18 é possível visualizar, respectivamente, a comparação entre os resultados obtidos pelos dicionários OpLexicon, SentiLex e SentiWordNet antes e depois das modificações em relação a avaliação manual.

Figura 16 – OpLexicon antes e depois das modificações



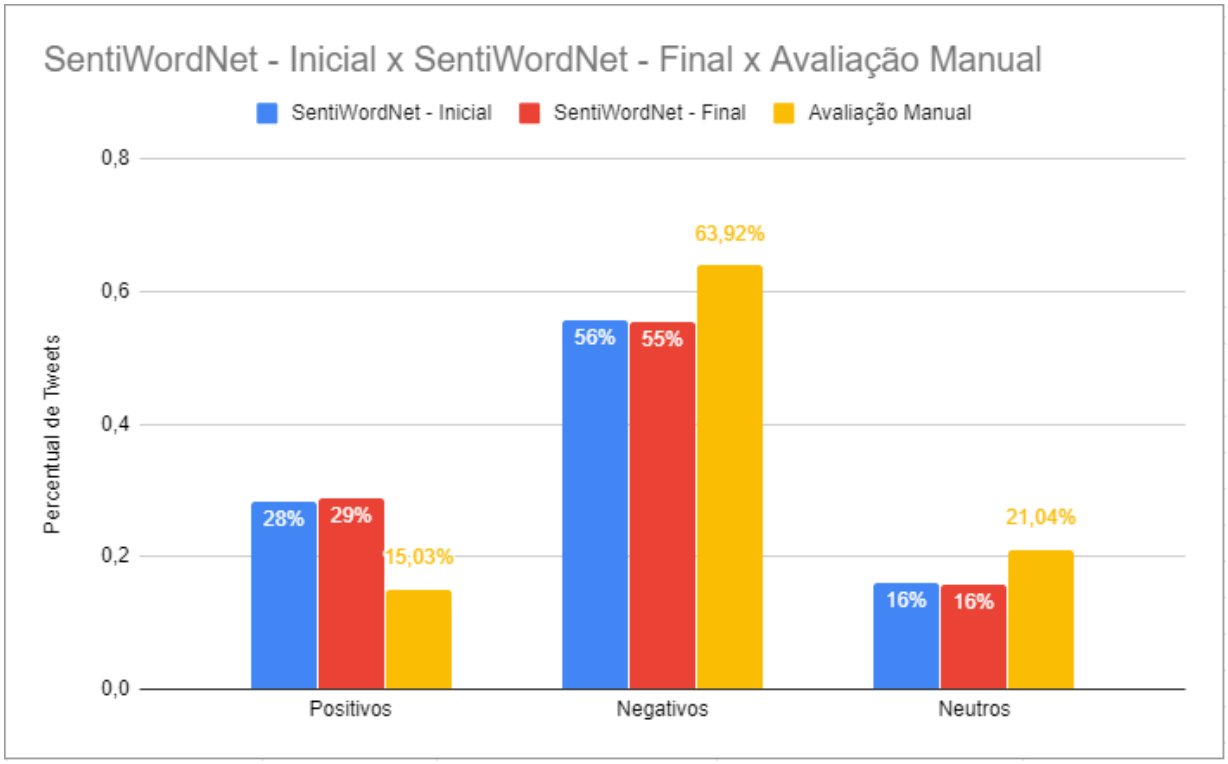
Fonte: Elaborada pelo autor

Figura 17 – SentiLex antes e depois das modificações



Fonte: Elaborada pelo autor

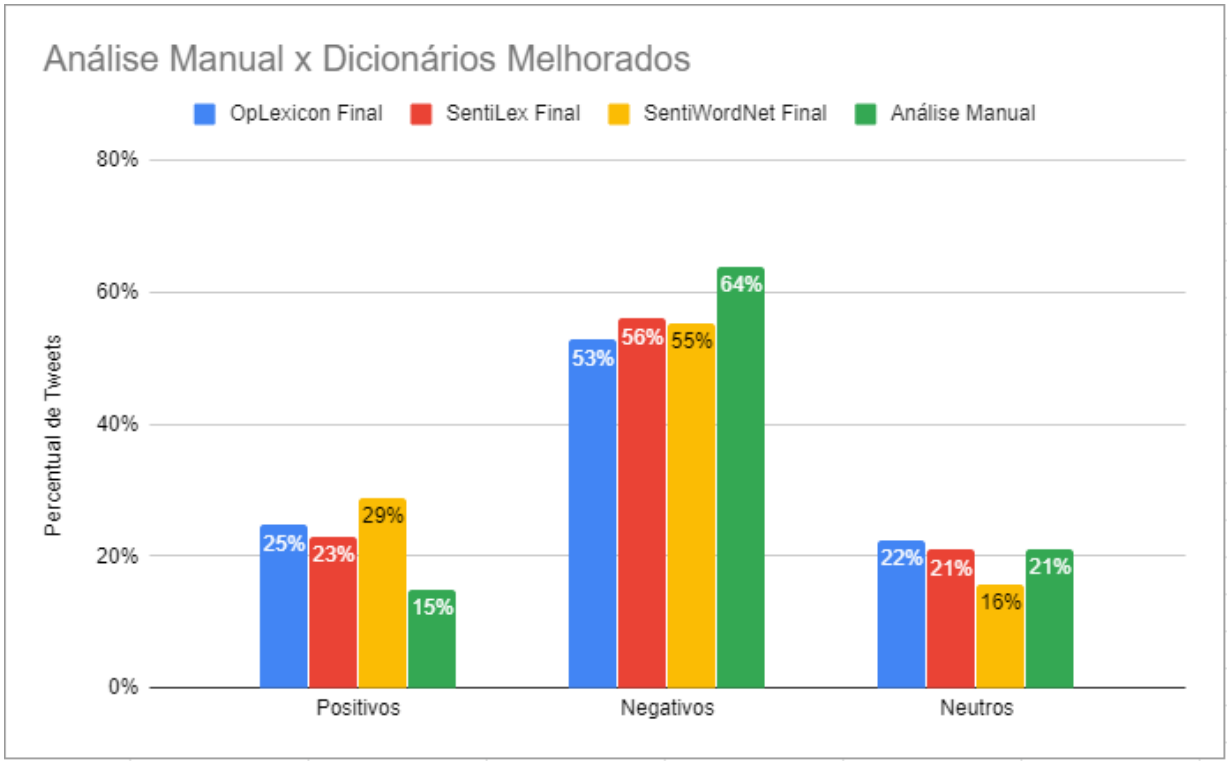
Figura 18 – SentiWordNet antes e depois das modificações



Fonte: Elaborada pelo autor

Ao serem feitas as melhorias nos dicionários, os resultados obtidos se aproximaram aos resultados da avaliação manual, além disso, com a diminuição da quantidade de *tweets* não processados, foi possível obter uma classificação mais correta dos *tweets*, assim como um resultado mais exato ao todo, pois com uma maior quantidade de *tweets* processados, mais exato seria o resultado final apresentado. Na Figura 19 temos a comparação da análise manual com a análise realiza através dos dicionários melhorados.

Figura 19 – Resultado final



Fonte: Elaborada pelo autor

Como pôde ser visto, a análise de alguns *tweets* pode ser dificultada por conta de palavras que são classificadas com um determinado valor mas, no contexto do *tweet*, pode adquirir um outro sentimento diferente do seu sentimento verdadeiro. Após a melhoria dos dicionários, com a adição de verbos no gerúndio, no pretérito perfeito e, a adição de palavras que estavam em grande número na base de dados, o valor do sentimento verdadeiro no *tweet* pôde ficar mais próximo de ser encontrado.

Para fins de comparação de *tweets* classificados corretamente e erroneamente, visualizaremos a seguir dois *tweets*, um classificado corretamente de acordo com os três dicionários léxicos e outro classificado de forma distinta pelos três dicionários, antes e depois das modificações. A Figura 20 representa um *tweet* que foi classificado corretamente pelos três dicionários.

Figura 20 – *Tweet* expressado de forma clara

Fonte: (TWITTER, 2019)

O dono da opinião na Figura 20 não expressou sua opinião de forma irônica nem fez uso de gírias ou quaisquer outros fenômenos linguísticos que poderiam levar a uma interpretação errônea do *tweet*. Ele se expressou de forma clara e objetiva, levando a uma classificação correta do *tweet* em todos os dicionários, antes mesmo de quaisquer modificações. Nas Figuras 21, 22 e 23 podemos ver, respectivamente, o *tweet* da Figura 20 sendo classificado corretamente pelos três dicionários léxicos.

Figura 21 – *Tweet* verdadeiro positivo - OpLexicon

```
>>> Tweet: 'pos o vídeo do Felipe Castanhari sobre as queimadas
na Amazônia ta genial demais, esse cara merece mais reconhecimento
Palavra Processada: genial
Valor Positivo: 1.0
Valor Negativo: 0.0
Valor do tweet: 1.0 - Positivo
```

Fonte: Elaborada pelo autor

Figura 22 – *Tweet* verdadeiro positivo - SentiLex[2]

```
>>> Tweet: 'pos o vídeo do Felipe Castanhari sobre as queimadas
na Amazônia ta genial demais, esse cara merece mais reconhecimento
Palavra Processada: genial
Valor Positivo: 1.0
Valor Negativo: 0.0
Valor do tweet: 1.0 - Positivo
```

Fonte: Elaborada pelo autor

[2]Tweet classificado da mesma forma que no dicionário OpLexicon, ilustrado na Figura 21

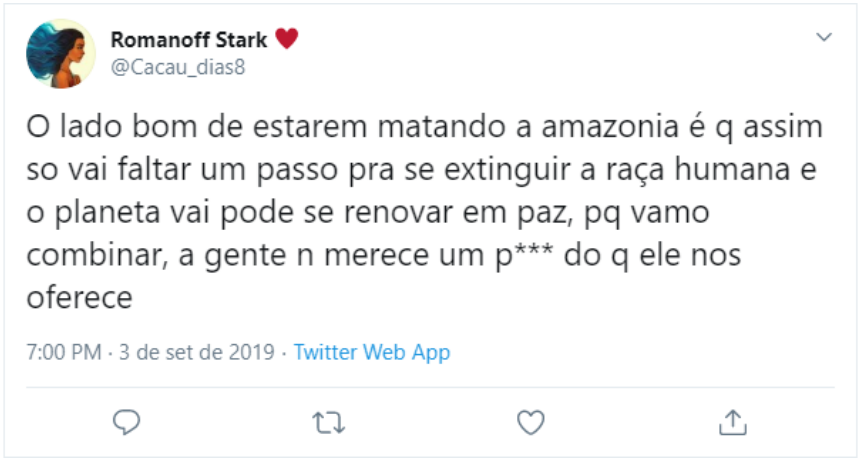
Figura 23 – *Tweet* verdadeiro positivo - SentiWordNet

```
>>> Tweet: 'pos o vídeo do Felipe Castanhari sobre as queimadas
na Amazônia ta genial demais, esse cara merece mais reconhecimento
Palavra Processada: demais
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: mais
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: Ás
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: cara
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: dó
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: O
Valor Positivo: 0.0
Valor Negativo: 0.125
Valor do tweet: 0.125 - Positivo
```

Fonte: Elaborada pelo autor

Na Figura 24, veremos um *tweet* que, antes das melhorias realizadas nos dicionários, foi classificado erroneamente pelo dicionário *SentiLex*.

Figura 24 – *Tweet* Negativo[3]



Fonte: (TWITTER, 2019)

Ao ser classificado pelo dicionário *SentiLex* antes das melhorias, o *tweet* da Figura 24 foi classificado como positivo, como mostrado na Figura 25. Observamos que esta classificação foi feita de forma errônea, pois o *tweet* tem um valor de sentimento negativo.

[3]Tweet censurado

Figura 25 – *Tweet* classificado erroneamente[4]

```
>>> Tweet: 'neg O lado bom de estarem matando a amazonia e q
assim so vai faltar um passo pra se extinguir a raca
humana e o planeta vai pode se renovar em paz,
pq vamo combinar, a gente n merece um p*** do q ele nos oferece
Palavra Processada: humana
Valor Positivo: 1.0
Valor Negativo: 0.0
Palavra Processada: bom
Valor Positivo: 1.0
Valor Negativo: 0.0
Palavra Processada: p***
Valor Positivo: 0.0
Valor Negativo: -1.0
Valor do tweet: 1.0 - Positivo
```

Fonte: Elaborada pelo autor

No *tweet* acima, podemos observar que o dono da opinião não expressou seu sentimento de maneira tão clara quanto o dono da opinião do *tweet* representado na Figura 20. O *tweet* contém palavras com sentido positivo, como a palavra "bom" e "humana", que fazem com que o sentimento total do *tweet* seja classificado como positivo. A ausência de palavras no dicionário que poderiam ajudar a diminuir o valor total do *tweet*, tornando-o negativo e, a falta de clareza do mesmo, contribuem para que o *tweet* seja classificado de maneira errônea.

Na Figura 26, o mesmo *tweet*, após as modificações, passa a apresentar o seu valor verdadeiro, negativo. Devido a adição de novas palavras no dicionário, algumas palavras antes ignoradas, que fariam diferença no sentido valor total da frase, como a palavra "matando", teve sua pontuação atribuída, contribuindo para a formação da pontuação correta do *tweet*.

[4] Avaliação censurada

Figura 26 – *Tweet* classificado corretamente[5]

```
>>> Tweet: 'neg O lado bom de estarem matando a amazonia e q
assim so vai faltar um passo pra se extinguir a raca
humana e o planeta vai pode se renovar em paz,
pq vamo combinar, a gente n merece um p*** do q ele nos oferece
Palavra Processada: humana
Valor Positivo: 1.0
Valor Negativo: 0.0
Palavra Processada: p***
Valor Positivo: 0.0
Valor Negativo: -1.0
Palavra Processada: matando
Valor Positivo: 0.0
Valor Negativo: -1.0
Palavra Processada: so
Valor Positivo: 0.0
Valor Negativo: -1.0
Palavra Processada: se
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: bom
Valor Positivo: 1.0
Valor Negativo: 0.0
Palavra Processada: pq
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: se
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: vai
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: n
Valor Positivo: 0.0
Valor Negativo: -1.0
Palavra Processada: gente
Valor Positivo: 0.0
Valor Negativo: 0.0
Palavra Processada: vai
Valor Positivo: 0.0
Valor Negativo: 0.0
Valor do tweet: -2.0 - Negativo
```

Fonte: Elaborada pelo autor

Como resultado final após as modificações realizadas nos dicionários, podemos observar que houve uma melhora na classificação dos *tweets* em relação aos resultados anteriores a melhoria. O *SentiLex* foi dicionário léxico que teve os seus resultados mais próximos a análise manual. O percentual de *tweets* classificados como neutros pelo *SentiLex* após as melhorias foi o mesmo percentual obtido na avaliação manual da base. Houve também um acréscimo no percentual de *tweets* negativos, aproximando-se da quantidade avaliada manualmente. Assim como o percentual de positivos, em que houve uma queda sutil, se aproximando a avaliação manual. No dicionário *OpLexicon* também houve uma aproximação dos resultados em todos os sentimentos, porém menor que no dicionário

[5] Avaliação censurada

SentiLex. Já no dicionário *SentiWordNet* não houve tantas mudanças significativas, assim, manteve um resultado destoante dos demais.

3.2.5 Comparação com Trabalhos Relacionados

Uma análise de sentimentos pode ser realizada de diversas maneiras, assim como pode-se utilizar uma ou mais abordagens. Por conta disso, a variedade de metodologias produzidas durante uma análise de sentimentos é comum devido a grande quantidade de recursos. Em razão da grande quantidade de metodologias possíveis, é comum que haja divergências durante a comparação entre trabalhos relacionados.

Dos trabalhos apresentados no Capítulo 1, todos utilizaram a mesma técnica e nível de análise de sentimentos empregados neste trabalho. No trabalho de Kolchyna et al. (2015), não só foi utilizada a abordagem léxica, como também a técnica de aprendizado de máquina, assim como no trabalho de Evangelista e Padilha (2014).

No trabalho de Costa (2017) foi utilizado apenas um dicionário léxico, o *SentiWordNet*, também utilizado neste trabalho, mas na versão de língua portuguesa. A semelhança em relação a esta monografia foi o uso do paralelismo para o processamento dos dados, fazendo o uso do *framework* Apache Ignite e do modelo de programação MapReduce, ambos obtendo um ótimo resultado em relação a velocidade do processo. O trabalho contou com 94 *tweets* na sua base de dados. A avaliação manual do autor classificou 62 comentários como positivos e 32 como negativos, enquanto que a avaliação produzida pelo dicionário classificou 63 comentários como positivos e 31 como negativos. Na melhor classificação deste trabalho, 23% dos *tweets* foram classificados como positivos e 56% como negativos, comparados a 15% classificados como positivos e 64% como negativos na avaliação manual. Costa (2017) obteve um ótimo resultado, porém, devido a pouca quantidade de *tweets* presentes na base, em comparação com este trabalho, que contém 2481 *tweets*, é válido que haja questionamento dos resultados.

A quantidade de *tweets* analisados por Evangelista e Padilha (2014) foi de 21 comentários no experimento realizado com a empresa Y, 13 no experimento com a empresa X e 38 no experimento com a empresa Z, em comparação com 2481 *tweets* classificados neste monografia. Devido a pouca quantidade de *tweets* analisados, não é possível retirar resultados satisfatórios. O trabalho obteve uma baixa taxa de acertos, com uma média de, aproximadamente, 53% de acertos dentre os *tweets* avaliados a respeito de três empresas. Além disso, pelo fato dos *tweets* capturados serem de contas comerciais, grande parte dos *tweets* continham propagandas, tendo assim, um sentimento neutro. Apesar disso, o dicionário utilizado quase sempre classificava uma palavra do *tweet* como positivo ou negativo, atribuindo um sentimento diferente de neutro para um *tweet* que, previamente, já era conhecido como neutro. Neste trabalho, não houveram problemas em classificar *tweets*, pois, a maioria dos *tweets* capturados continham opiniões pessoais sobre um assunto

polêmico, podendo assim, obter sentenças com diferentes tipos de sentimento.

No trabalho de Kolchyna et al. (2015), foram utilizadas diferentes formas de classificação, com a utilização de dois cálculos de polaridades distintos. A configuração de dicionários que obteve o melhor resultado excluiu a utilização do dicionário automatizado, pois, durante o treinamento para a criação do dicionário, os ruídos dos dados extraídos do Twitter acabaram gerando ambiguidade nas palavras extraídas, obtendo um total de 52,38% de *tweets* classificados corretamente com o melhor cálculo de polaridade. O dicionário de *emoticons* contribuiu para resultados melhores no trabalho do autor. Com a melhor configuração empregada no trabalho, fazendo uso apenas do dicionário léxico e do dicionário de *emoticons*, Kolchyna et al. (2015) obteve 61,74% dos *tweets* classificados corretamente, demonstrando que nem sempre uma abordagem automática pode ser eficaz se tratando de um grande conjunto de dados com uso de abreviações e palavras escritas de maneira errada, como é o caso do Twitter.

Este capítulo apresentou as etapas, características e desafios da ferramenta desenvolvida para o estudo de caso. Também foram destacados os resultados obtidos e conclusões acerca dos mesmos.

No próximo capítulo serão feitas as conclusões acerca deste trabalho, assim como sugestões de trabalhos futuros.

4 Conclusão

Este trabalho teve como objetivo produzir uma comparação de resultados produzidos por diferentes dicionários léxicos a partir de uma metodologia proposta em uma análise de sentimentos a nível de sentença. Essa análise foi produzida em uma base de *tweets* relacionados a uma série de focos de incêndio que ocorreram na região amazônica do Brasil no mês de setembro de 2019.

O desenvolvimento deste trabalho foi realizado com o uso do *framework* Apache Ignite. O *framework* se mostrou a melhor opção em comparação ao Apache Hadoop devido às suas características introduzidas no tópico 2.2.3.

A metodologia empregada neste trabalho foi executada por meio de três dicionários léxicos. Após os resultados, foi constatado que a análise de sentimentos a nível de sentença não é uma tarefa simples, pois, principalmente em redes sociais, há muitos erros ortográficos e ambiguidades, assim como o uso de gírias e abreviações, o que dificultou a produção dos resultados. Após a análise, foram realizadas um conjunto de melhorias em cada um dos dicionários, para que pudessem ser obtidos resultados com uma maior acurácia em comparação com uma avaliação manual dos *tweets* realizada pelo autor do trabalho. Após as modificações realizadas nos dicionários, houve uma melhora na classificação, obtendo-se um resultado excelente na utilização de todos os dicionários, principalmente com o uso do dicionário *SentiLex*.

Apesar dos desafios encontrados, os resultados da análise, em geral, foram satisfatórios, principalmente após as modificações nos dicionários. Além disso, o uso do *framework* Apache Ignite, o modelo de programação MapReduce também se mostrou útil durante o desenvolvimento do trabalho, reduzindo o tempo de execução consideravelmente após ser feito o uso do paralelismo da execução.

Em trabalhos futuros, pretende-se adaptar os dicionários para que possam ser capazes de detectar variações linguísticas vistas durante a análise, assim como gírias e abreviações. Além disso, pretende-se fazer o uso de um dicionário de *emoticons*, como empregado por (KOLCHYNA et al., 2015). Também propõe-se incluir na metodologia a possibilidade de uma análise por meio do emprego de ferramentas de análise automática, utilizando uma abordagem híbrida na metodologia, preocupando-se com os tipos de dados a serem tratados, para que a abordagem se adéque ao tipo de conteúdo analisado. Por último, planeja-se uma análise mais detalhada dos resultados, propondo o resultado de *tweets* analisados corretamente e erroneamente de maneira individual, em todos os dicionários, antes e depois das modificações.

Referências

APACHE HADOOP. *Documentation*. 2019. Disponível em: <https://hadoop.apache.org/docs/stable/>. Acesso em: 26 nov. 2019. Citado 2 vezes nas páginas 21 e 22.

APACHE IGNITE. *Documentation*. 2019. Disponível em: <https://apacheignite.readme.io/docs/what-is-ignite>. Acesso em: 19 nov. 2019. Citado 3 vezes nas páginas 23, 24 e 30.

BEYER, M.; LANEY, D. *The Importance of 'Big Data': A Definition*. 2012. Citado 2 vezes nas páginas 13 e 17.

CAVANILLAS, J. M. et al. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. [S.l.]: SpringerBriefs in Computer Science, 2016. Citado na página 19.

CHEN, M. et al. *Big Data Related Technologies, Challenges and Future Prospects*. [S.l.]: SpringerBriefs in Computer Science, 2014. Citado 3 vezes nas páginas 13, 17 e 19.

COSTA, A. C. B. *Análise de sentimentos em nível de sentença a partir de dados extraídos do Twitter utilizando o framework Apache Ignite*. [S.l.]: Universidade Federal do Maranhão, 2017. Disponível em: <https://goo.gl/i0pRB>. Acesso em: 23 jun. 2019. Citado 2 vezes nas páginas 15 e 50.

DEAN, J.; GHEMAWAT, S. *MapReduce: Simplified Data Processing on Large Clusters*. 2014. Disponível em: <https://static.googleusercontent.com/media/research.google.com/pt-BR//archive/mapreduce-osdi04.pdf>. Acesso em: 20 nov. 2019. Citado 2 vezes nas páginas 20 e 21.

DUBEY, V.; GUPTA, D. L. Sentiment analysis using singular value decomposition. *International Journal of Current Engineering and Technology*, v. 6, n. 4, ago, 2016. Acesso em: 11 dez. 2019. Citado na página 26.

ECLIPSE FOUNDATION. *Eclipse Foundation*. 2019. Disponível em: <https://www.eclipse.org/org/foundation/>. Acesso em: 19 nov. 2019. Citado na página 30.

EVANGELISTA, T. R.; PADILHA, T. P. P. *Monitoramento de Posts Sobre Empresas de E-Commerce em Redes Sociais Utilizando Análise de Sentimentos*. 2014. Citado 2 vezes nas páginas 15 e 50.

FUNÇÃO SISTEMAS. *Big Data*. 2019. Disponível em: <https://www.funcao.com.br/2019/01/18/big-data/>. Acesso em: 26 nov. 2019. Citado na página 18.

GADOMI, A.; HAIDER, M. *Beyond the hype: Big data concepts, methods, and analytics*. 2015. Citado na página 17.

KOLCHYNA, O. et al. *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*. 2015. Disponível em: <https://arxiv.org/abs/1507.00955>. Acesso em: 23 jun. 2019. Citado 4 vezes nas páginas 15, 50, 51 e 52.

KOLKUR, S.; DANTAL, G.; MAHE, R. *Study of Different Levels for Sentiment Analysis*. 2015. Citado na página 26.

LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan Claypool Publishers, 2012. Citado 4 vezes nas páginas 13, 24, 25 e 27.

MOREIRA, V. de S. et al. *Análise de Sentimentos: Comparando o uso de ferramentas e a análise humana*. 2016. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2016/058.pdf>. Acesso em: 23 jun. 2019. Citado na página 14.

OPLEXICON. 2019. Disponível em: <https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/>. Acesso em: 03 nov. 2019. Citado na página 34.

PANG, B.; LEE, L. *Opinion mining and sentiment analysis*. 2008. Foundations and Trends in Information Retrieval. Citado 2 vezes nas páginas 27 e 28.

PANG, B. et al. *Thumbs up? Sentiment Classification using Machine Learning Techniques*. 2002. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Citado na página 27.

REVISTA PLANETA. *Metade dos habitantes do planeta está nas redes sociais*. 2019. Disponível em: <https://www.revistaplaneta.com.br/metade-dos-habitantes-do-planeta-esta-nas-redes-sociais/>. Acesso em: 18 nov. 2019. Citado na página 13.

REVISTA PLANETA. *Sentiment analysis: dados de sentimentos geram insights poderosos*. 2019. Disponível em: <https://www.revistaplaneta.com.br/metade-dos-habitantes-do-planeta-esta-nas-redes-sociais/>. Acesso em: 26 nov. 2019. Citado na página 13.

RUIZ, D. H. F. e E. E. S. *Análise de Sentimentos aplicada à realidade da doação de sangue no Brasil usando dados do Twitter*. 2016. Disponível em: <http://docs.bvsalud.org/biblioref/2018/07/906570/anais_cbis_2016_artigos_completos-653-660.pdf>. Acesso em: 08 nov. 2019. Citado na página 13.

SENTILEX. 2019. Disponível em: <http://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3>. Acesso em: 19 nov. 2019. Citado na página 34.

SENTIWORDNET-PT-BR. 2019. Disponível em: <https://github.com/Pedro-Thales/SentiWordNet-PT-BR>. Acesso em: 03 nov. 2019. Citado na página 34.

SILVA, L. L. A. A. *Análise de sentimentos em contexto: estudo de caso em blog de empreendedorismo*. 2013. Instituto de Ciências Exatas Departamento de Ciência da Computação. Citado na página 26.

TWEEPY. *An easy-to-use Python library for accessing the Twitter API*. 2019. Disponível em: <https://www.tweepy.org/>. Acesso em: 03 nov. 2019. Citado na página 30.

TWITTER. *Twitter*. 2019. Disponível em: <https://twitter.com/home>. Acesso em: 08 nov. 2019. Citado 2 vezes nas páginas 46 e 47.

TWITTER DEVELOPER. *Twitter Developer*. 2019. Disponível em: <https://developer.twitter.com/>. Acesso em: 19 nov. 2019. Citado na página 31.

WIEBE, J. M. *Development and use of a gold-standard data set for subjectivity classifications*. 1999. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). Citado na página 27.