

**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CIÊNCIA DA COMPUTAÇÃO**

FILIPPE RODRIGUES PEREIRA DA SILVA

**TSCOOP - AVALIAÇÃO DE FERRAMENTAS PARA ANÁLISE DE
SENTIMENTO: UM ESTUDO DE CASO NO TWITTER**

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA

2019

FILIFE RODRIGUES PEREIRA DA SILVA

**TSCOOP - AVALIAÇÃO DE FERRAMENTAS PARA ANÁLISE DE
SENTIMENTO: UM ESTUDO DE CASO NO TWITTER**

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel, Ciência da Computação, do Departamento Acadêmico de Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Richard Duarte Ribeiro

Coorientadora: Profa. Dra. Helyane Bronoski Borges

PONTA GROSSA

2019



Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Ponta Grossa

Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Bacharelado em Ciência da Computação



TERMO DE APROVAÇÃO

TSCOOP – AVALIAÇÃO DE FERRAMENTAS DE ANÁLISE DE SENTIMENTO: UM ESTUDO DE CASO NO TWITTER

por

FILIPE RODRIGUES PEREIRA DA SILVA

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 05 de junho de 2019 como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Dr. Richard Duarte Ribeiro
Orientador

Profa. Dra. Helyane Bronoski Borges
Co-orientadora

Profa. Dra. Simone de Almeida
Membro titular

Profa. Dra. Eliana Cláudia Mayumi Ishikawa
Membro titular

Prof. MSc. Geraldo Ranthum
Responsável pelo Trabalho de Conclusão
de Curso

Prof. MSc. Saulo Jorge Beltrão de Queiroz
Coordenador do curso

Dedico este trabalho a todos os meus
familiares que já não estão mais entre
nós.

AGRADECIMENTOS

Agradeço ao meu orientador Richard pela paciência, sabedoria e por sempre acreditar em minha capacidade, mesmo em momentos que eu mesmo não fui capaz.

Agradeço a minha família pela oportunidade de vivenciar tudo isso.

Agradeço aos meus amigos que me dão diariamente motivação para continuar e pelo apoio nesse caminho.

Agradeço a minha coorientadora Helyane, cujas sugestões puderam direcionar o objetivo deste trabalho e como isso pode se tornar algo maior.

“Love is bigger than anything in its way” – U2

RESUMO

SILVA, Filipe R. P. **Tscoop – Avaliação de ferramentas de análise de sentimento: um estudo de caso no Twitter**. 2019. 46f. Trabalho de Conclusão de Curso Superior de Bacharelado em Ciência da Computação - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2019.

Atualmente vivemos um momento em que as redes sociais fazem parte da rotina de muitas pessoas ao redor do mundo. Benefícios como o acesso facilitado de informação pode se tornar algo negativo pelo fato de existirem pessoas que pretendem utilizar redes sociais como ferramenta para manipular e influenciar opiniões. Utilizando a rede social Twitter, pretende-se avaliar sentimentos sobre *tweets* de um determinado assunto, com o auxílio de ferramentas de análise de sentimentos (Microsoft Azure, Repustate e Meaning Cloud), que também tiveram o desempenho avaliado.

Palavras-chave: Twitter. Análise de Sentimento. Redes Sociais.

ABSTRACT

SILVA, Filipe R. P. **Tscoop - Evaluation of sentiment analysis tools: a case study on Twitter**. 2019. 46p. Conclusion of the Bachelor's Degree in Computer Science - Federal Technological University of Paraná. Ponta Grossa, 2019.

Today we are living in a time when social networks are part of the routine of many people around the world. Benefits such as facilitating access to information can become somewhat negative because there are people who want to use social networks as a tool to manipulate and influence opinions. Using the social network Twitter, it is intended to evaluate sentiments about *tweets* of a certain subject, with the aid of tools of sentiment analysis (Microsoft Azure, Repustate and Meaning Cloud) that will also have their performance evaluated.

Keywords: Twitter. Sentiment Analysis. Social Networks.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de uma url para buscar tweets que possuam a palavra “nasa” entre os tweets mais populares.....	20
Figura 2 – Trecho de um tweet resultado da busca feita anteriormente.....	20
Figura 3 – Passos da mineração de texto	21
Figura 4 – Diagrama generalizado da análise de sentimento	24
Figura 5 – Estrutura generalizada da ferramenta	30
Figura 6 – Migration da tabela tweets	31
Figura 7 – Código da model Tweet	32
Figura 8 – Modelo relacional do banco de dados.....	33
Figura 9 – Query de busca na API do Twitter.....	34
Figura 10 – Laço principal de iteração sobre o conjunto de Tweets.....	34
Figura 11 – Verificação da existência da palavra de busca e armazenamento do tweet.....	35
Figura 12 – Código do método parseTweet().....	35
Figura 13 – Código do método analyzeSentence()	36
Figura 14 – Código do método analyzeSentenceAzure().....	36
Figura 15 – Código do método sentiment()	37
Figura 16 – Código do método storeSentiment().....	37
Figura 17 – Diagrama de execução do Tscoopws	38
Quadro 1 – Relação da pontuação e sentimento – Microsoft Azure	22
Quadro 2 – Relação da pontuação e sentimento – Repustate	23
Quadro 3 – Relação da pontuação e sentimento – Meaning Cloud	23
Quadro 4 – Avaliação de sentimento das ferramentas em relação ao tweet 1	44
Quadro 5 – Avaliação de sentimento das ferramentas em relação ao tweet 2	44
Gráfico 1 – Gráfico demonstrativo do tempo de uso diário de aplicativos.....	14
Gráfico 2 – Gráfico do estado sentimental dos dias 25 e 26/03.....	39
Gráfico 3 – Média da avaliação de sentimento de todas as ferramentas nos dias 19 e 20/04	40
Gráfico 4 – Avaliação de Sentimento da ferramenta Repustate nos dias 19 e 20/04.....	41
Gráfico 5 – Avaliação de Sentimento da ferramenta Microsoft Azure nos dias 19 e 20/04	41
Gráfico 6 – Avaliação de Sentimento da ferramenta Meaning Cloud nos dias 19 e 20/04	41
Gráfico 7 – Média da avaliação de sentimento de todas as ferramentas nos dias 12 e 13/05	42
Gráfico 8 – Avaliação de Sentimento da ferramenta Microsoft Azure nos dias 12 e 13/05	42
Gráfico 9 – Avaliação de Sentimento da ferramenta Meaning Cloud nos dias 12 e 13/05	43

Gráfico 10 – Avaliação de Sentimento da ferramenta Repustate nos dias 12 e 13/05
.....43

SUMÁRIO

1 INTRODUÇÃO	13
1.1 OBJETIVOS	15
1.1.1 Objetivo Geral	15
1.1.2 Objetivos Específicos	15
1.2 CONTRIBUIÇÕES DESTE TRABALHO	16
1.3 MOTIVAÇÃO	16
1.4 ESTRUTURA DO DOCUMENTO	16
2 REFERENCIAL TEÓRICO	18
2.1 TWITTER	18
2.2.1 TWITTER API	19
2.2.1.1 API de busca padrão	19
2.2 MINERAÇÃO DE TEXTO	21
2.3 ANÁLISE DE SENTIMENTO	22
2.4 FERRAMENTAS DE ANÁLISE DE SENTIMENTO	23
2.4.1 Microsoft Azure – <i>Text Analytics</i>	24
2.4.2 Repustate – <i>Sentiment Analysis</i>	25
2.4.3 Meaning Cloud – <i>Sentiment Analysis</i>	26
2.5 CONSIDERAÇÕES DO CAPÍTULO	27
3 ESTUDO DE CASO	28
3.2.1 Estrutura	29
3.2.1.1 Migrations	30
3.2.1.2 Models	31
3.2.1.3 PostgreSQL	32
3.2.2 Busca e pré-processamento	33
3.2.3 Análise de sentimento	35
3.3 RESULTADOS	38
4 CONCLUSÃO	45
4.1 TRABALHOS FUTUROS	45
REFERÊNCIAS	47

1 INTRODUÇÃO

De acordo com uma pesquisa publicada pelo *website* Facebook em 2016, estima-se que 3,2 bilhões de pessoas já possuíam acesso à Internet em 2015, totalizando 43% da população mundial conectadas na rede (FACEBOOK, 2016). Adicionalmente, segundo dados publicados em uma pesquisa do Centro Gestor da Internet no Brasil (CGIBR) em 2015, 102 milhões de brasileiros possuíam acesso à Internet, totalizando 58% da população brasileira conectada (CGI, 2016).

Em paralelo com o crescimento da Internet, os meios de acessá-la também cresceram e novas tecnologias foram desenvolvidas para os telefones celulares e computadores. Como exemplo dessa evolução pode-se notar na pesquisa realizada pela CGIBR (2016) que 89% dos usuários utilizavam a Internet pelo telefone celular, tornando o celular uma ferramenta essencial na vida de muitas pessoas, seja para trabalho e/ou lazer.

Combinando a baixa nos preços de planos para Internet móvel (FACEBOOK, 2018) e a presença massiva dos aparelhos celulares na vida das pessoas, o acesso às redes sociais se tornou ainda mais fácil para os diversos tipos de usuários.

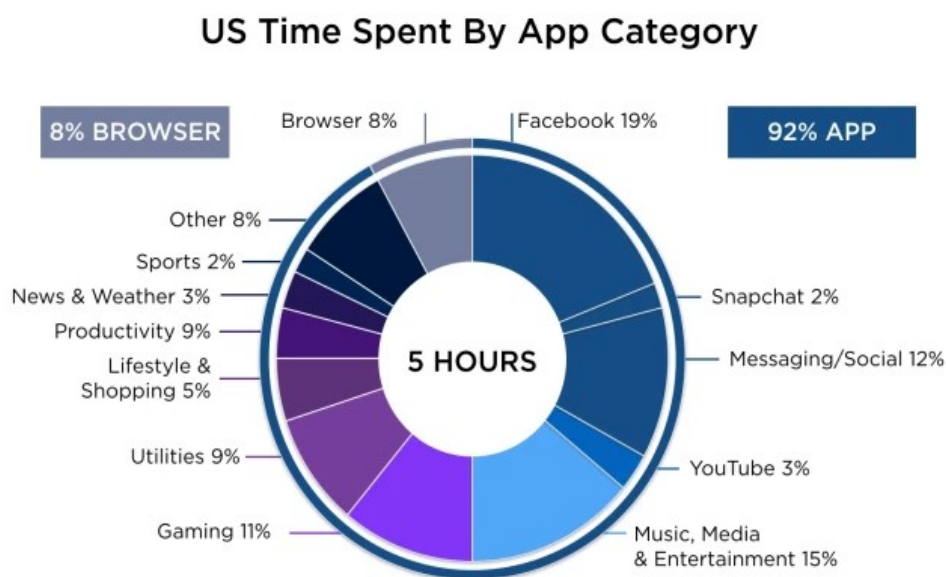
Devido à grande presença da Internet e dos aparelhos eletrônicos no dia a dia, as pessoas podem se tornar usuárias compulsivas dessa tecnologia, conseqüentemente aumentando sua exposição a diversos riscos inerentes ao seu uso. Um desses riscos está relacionado ao uso intensivo de redes sociais, que pode gerar exposição excessiva a propagandas direcionadas e, no caso da rede social Twitter, exposição a diversos *Tweets* falsos e/ou até mesmo manipulados para criar falsas tendências (BBC, 2018).

Pessoas com más intenções podem tirar proveito desse uso regular das redes sociais para influenciar os usuários a cometer diversos tipos de ações, como votar em um candidato X ou Y, comprar determinado tipo de produto ou até mesmo radicalizar opiniões políticas.

Nota-se pelo Gráfico 1 publicado pela empresa Flurry Analytics (2017), que os cidadãos dos Estados Unidos gastam em média 5 horas diárias em seus telefones celulares. A exposição exagerada às redes sociais pode tornar os usuários suscetíveis a diversas formas de influência devido ao tempo regularmente gasto nas mesmas.

Um caso recente de mal-uso das propagandas no Facebook está relacionado à Rússia, que supostamente conspirou para influenciar a eleição presidencial de 2016 nos Estados Unidos em favor de Donald Trump e contra Hillary Clinton (THE GUARDIAN, 2018).

Gráfico 1 – Gráfico demonstrativo do tempo de uso diário



Source: Flurry Analytics, comScore, Facebook, NetMarketShare. Note: US, Dec 2016

Fonte: Flurry Analytics (2017).

O Departamento de Justiça dos Estados Unidos anunciou em fevereiro de 2018 que auxiliados pelo ex-diretor do FBI, Robert Mueller indiciaram 13 russos e três entidades russas em uma suposta conspiração para fraudar os Estados Unidos. A acusação mostra que houve um planejamento muito bem detalhado durante 3 anos para influenciar, incitar discórdia política e causar danos a campanha presidencial de Hillary Clinton (DOJ, 2018).

Segundo uma investigação da BBC (2018), existem empresas na Arábia Saudita que estão oferecendo serviços para modificar artificialmente os tópicos mais comentados do Twitter no país por algumas horas. Essa prática pode servir para promover uma determinada marca, produto ou até mesmo para espalhar diversas notícias falsas na rede social.

Ainda no relatório do primeiro quarto de 2018, o Twitter detectou mais de 142 mil aplicativos conectados a API (*Application Programming Interface* - Interface de Programação de Aplicação) do Twitter que estavam violando as políticas de uso, produzindo mais de 130 milhões de *Tweets* considerados provenientes de contas falsas (TWITTER, 2018).

Considerando a utilização massiva das redes sociais atualmente, um estudo de influência se mostra promissor. Uma possível utilização de análise deste gênero é observar como os sentimentos das pessoas mudam conforme assuntos diversos começam a surgir nas redes sociais. A pergunta a ser respondida é: como mudou o sentimento das pessoas em relação a uma determinada pessoa após uma *fake news* ou um vazamento?

Para que seja possível responder essa pergunta, se faz necessário o desenvolvimento de uma ferramenta de análise de sentimentos.

Os estudos levantados até o momento não apresentaram uma forma possível de se fazer isso. A seguir, será apresentado os objetivos que se alinham com o possível problema de influências que estão sendo exploradas atualmente.

1.1 OBJETIVOS

O objetivo geral e os específicos são descritos a seguir.

1.1.1 Objetivo Geral

Desenvolver uma ferramenta que estude os sentimentos em textos relacionados a um tema específico.

1.1.2 Objetivos Específicos

Desenvolver um *webservice* capaz de manter um fluxo contínuo com a API do Twitter para transmissão de dados.

Submeter os dados obtidos no Twitter nas ferramentas de análise de sentimento: *Microsoft Azure, Repustate e Meaning Cloud*.

Avaliar o desempenho individual das ferramentas utilizadas para realizar a análise de sentimento.

Estudar e comparar os resultados obtidos das ferramentas utilizadas.

1.2 CONTRIBUIÇÕES DESTE TRABALHO

Este trabalho contribui com a criação da ferramenta TScoopws¹ para integrar as ferramentas de análise de sentimento e de busca por *tweets*. Foi possível analisar as mudanças de sentimento que são visualizadas por meio de gráficos e as diferenças entre as classificações das ferramentas. No restante do trabalho as fases de desenvolvimento são detalhadas.

1.3 MOTIVAÇÃO

Após o vazamento dos dados de mais de 87 milhões de usuários do Facebook para a empresa Cambridge Analytica, de acordo com Forbes (2018), é possível que diversas empresas estejam utilizando os perfis de usuários para criar propagandas direcionadas para atingir os usuários de diversas redes sociais.

Por meio da detecção de perfis de usuários em determinada região, é possível que haja uma tentativa de falsificação dos tópicos mais comentados no Twitter para tentar influenciar usuários a comentarem sobre um determinado assunto.

Desenvolver uma ferramenta capaz de identificar o estado e o histórico sentimental da rede pode ter alto valor para estudos futuros envolvendo o Twitter e análise de sentimento.

1.4 ESTRUTURA DO DOCUMENTO

O referencial teórico contém informações sobre como os dados que influenciaram este trabalho foram levantados, além de apresentar pontos que apoiam a ideia de influência, brevemente citada nesta seção introdutória.

¹ Este trabalho pode ser encontrado em <https://bitbucket.org/FilipeRPSilva/tscoopws>.

Detalhes das ferramentas de análise de sentimento se encontram na seção 2.3. A seção 3.2 apresenta como o trabalho foi desenvolvido, desde as tecnologias usadas, modo de utilização (Seção 3.3) e os resultados obtidos (Seção 3.4).

No último capítulo, está descrito de forma mais detalhada as contribuições deste projeto, além de as perguntas que foram deixadas em abertas como sugestões a trabalhos futuros.

2 REFERENCIAL TEÓRICO

Nesta seção são definidos os componentes essenciais para embasamento teórico do trabalho. Começa-se explicando o que é mineração de dados, seguindo de um detalhamento sobre a análise de sentimento, finalizando com as ferramentas selecionadas para análise.

2.1 TWITTER

O Twitter é uma rede social que permite aos usuários publicar e ver atualizações de outros usuários por meio dos chamados *tweets*, que são textos com no máximo 280 caracteres. Normalmente os *tweets* estão associados a algum tópico, representado por um símbolo de *hashtag* “#” seguido do assunto, por exemplo “#utfpr”.

Quanto mais vezes um mesmo tópico for publicado por diferentes usuários, mais chances esse tópico terá de atingir a lista de tópicos populares, que pode ser agrupado por *tweets* do mundo inteiro, países e algumas cidades. A lista de tópicos mais comentados pode servir para a identificação do alcance de um evento, notícia ou até mesmo o acontecimento de um fenômeno natural.

Para poder ver e acompanhar as atualizações de outro usuário, basta se tornar seguidor desse outro usuário. Quanto mais seguidores um usuário possuir, maior a chance que seus *tweets* sejam compartilhados por pessoas diferentes pela ferramenta de “*retweet*”.

Nos primeiro três meses de 2018, foram registrados aproximadamente 336 milhões de usuários mensalmente ativos na rede social Twitter, gerando diariamente mais de 500 milhões de *tweets* por dia (TWITTER, 2018). Também houve um crescimento de 10% do número de usuários diariamente ativos em relação aos últimos três meses de 2017.

Esses resultados mostram a importância dessa ferramenta. Adicionalmente, se analisados os resultados de anos anteriores, nota-se um constante crescimento da rede social (TWITTER, 2018). Tal crescimento é tão acelerado que houve um aumento na receita de aproximadamente 21% no Twitter (2018), que representa um resultado maior do que o esperado pela empresa.

2.2.1 Twitter API

A API² de serviços do Twitter fornece uma grande variedade de recursos para que seja possível realizar diversas ações sem que se esteja de fato conectado ao site ou aplicativo do Twitter.

A API de busca por *tweets* do Twitter está dividida em três categorias: padrão, premium e empresas: (A API padrão foi utilizada durante o desenvolvimento deste projeto)

- Padrão: Essa API gratuita é capaz de buscar em uma amostra dos *tweets* publicados nos últimos sete dias presentes no conjunto de *tweets* públicos.
- Premium: Acesso gratuito (na conta de desenvolvedor do twitter, que fornece acesso aos *tweets* dos últimos 30 dias) e acesso pago (para buscar *tweets* publicados desde 2006), em que a fidelidade dos dados é garantida, diferentemente da API padrão.
- Empresas: Acesso pago para acesso de todos os *tweets* publicados desde 2006, gerenciamento de conta e fidelidade total dos dados extraídos.

2.2.1.1 API de busca padrão

O principal objetivo da API de busca padrão é retornar uma coleção com os *tweets* mais relevantes baseado nos parâmetros de busca. A API de busca padrão funciona de maneira similar ao campo de busca padrão presente no *site* e aplicativo do Twitter.

Para utilizar a API de busca padrão é necessário que seja feita a autenticação via padrão OAuth e seja gerado o *token* de acesso. Após a autenticação, basta fazer uma requisição³. Os seguintes parâmetros podem ser utilizados para refinar as buscas: *q*, *geocode*, *lang*, *locale*, *result_type*, *count*, *until*, *since_id*, *max_id*. Onde cada parâmetro especificação é como se segue:

² <https://developer.twitter.com/>

³ <https://api.twitter.com/1.1/search/tweets.json>

- *q*: *string* de no máximo 500 caracteres contendo os termos que devem ser encontrados.
- *geocode*: restringe o retorno para os *tweets* de usuários localizados em um dado raio e latitude e longitude.
- *lang*: restringe o retorno para os *tweets* de uma determinada língua.
- *locale*: especifica a linguagem da busca contida no parâmetro “q”.
- *result_type*: especifica o tipo de resultado da busca, como os *tweets* mais recentes ou *tweets* mais populares.
- *count*: especifica o número de *tweets* a retornar por página.
- *until*: restringe os *tweets* criados até a data informada.
- *since_id*: retorna *tweets* com id maior que o informado (quanto maior o id, mais recente o *tweet*).
- *max_id*: retorna *tweets* com id menor que o informado.

Um exemplo de uma *url* de busca com alguns dos parâmetros citados está presente na Figura 3. Adicionalmente, pode-se ver na Figura 4 um pequeno trecho relacionado a um *tweet* proveniente do resultado esperado de uma requisição feita com a *url* da Figura 3.

Figura 1 – Exemplo de uma url para buscar tweets que possuam a palavra “nasa” entre os tweets mais populares.

```
twurl /1.1/search/tweets.json?q=nasa&result_type=popular
```

Fonte: Desenvolvedores Twitter (2018)

Figura 2 – Trecho de um tweet resultado da busca feita anteriormente.

```
"id" : 113482943,  
"id_str" : "113482943",  
"name" : "nasa",  
"screen_name" : "NASA",  
"location" : "",  
"description" : "Explore the universe and discover our home planet with @NASA."  
"url" : "https://t.co/TcEE6NS8nD"
```

Fonte: Desenvolvedores Twitter (2018)

2.2 MINERAÇÃO DE TEXTO

A mineração de textos pode ser definida como um método de extração de informações relevantes em bases de dados não estruturadas, ou semiestruturadas (FELDMAN e SANGER, 2006). Muitos aspectos da mineração de texto são semelhantes aos da mineração de dados, como: pré-processamento e descoberta de padrões (FELDMAN e SANGER, 2006).

A mineração de texto consiste em uma série de 5 etapas que podem ser realizadas em ordem para extrair informações de forma eficiente de um documento, sentença, texto etc. Na Figura 1 pode-se observar essas atividades e posteriormente entender o papel de cada um nesse processo.

Figura 3 - Passos da mineração de Texto



Fonte: Jiakang Chang, et al (2018).

A fase de coleta consiste em reunir dados de uma ou várias fontes como *websites*, e-mails, comentários em fóruns, arquivos, etc.(CHANG, et al, 2018).

A fase de pré-processamento é formada por três etapas principais: limpeza do texto, em que são removidos os caracteres ou trechos de texto indesejados. Tokenização, em que o texto é dividido em sentenças ou entidades, dependendo do objetivo da aplicação. Finalmente, a Seleção de Atributos faz a análise quantitativa do texto, podendo obter a frequência de uma determinada palavra no texto, por exemplo (CHANG, et al, 2018).

A fase de armazenamento consiste na criação de índices na base de dados e outros elementos que possibilitem o acesso mais rápido aos dados dentro de um banco de dados (CHANG, et al, 2018).

Na fase de mineração e análise ocorre a utilização de algoritmos e técnicas de exploração de dados para possivelmente revelar alguma informação de um texto.

Por fim, a fase de apresentação busca visualizar os dados para então avaliá-los e posteriormente chegar em alguma conclusão baseando-se em suas informações (CHANG, et al, 2018).

Uma aplicação da utilização da mineração de texto está em grandes empresas que trabalham com atendimento ao cliente, em que normalmente o cliente liga para a empresa para reclamar ou solicitar ajuda para um determinado produto e o cliente precisa entrar em contato com algum especialista da empresa para resolver seu problema. É comum ocorrer um tempo de espera muito grande pois a requisição do cliente é transferida de setor para setor, demorando muito para chegar no destino (CARRILHO JUNIOR, 2007).

Uma possível solução utilizando mineração de texto, seria o processamento da mensagem do cliente para identificar possíveis palavras chave de modo a facilitar e automatizar a transferência para o setor responsável, removendo a interação humana e agilizando o processo (CARRILHO JUNIOR, 2007).

2.3 ANÁLISE DE SENTIMENTO

A análise de sentimento se tornou um importante campo de pesquisa para entender a opinião das pessoas sobre um determinado assunto, produto, marca etc. analisando grandes quantidades de informações (SALINCA, 2015). O sentimento dessas opiniões não é valioso somente para as empresas analisarem a satisfação do cliente e monitorar seus concorrentes, mas pode ser valioso para o consumidor que desejar pesquisar sobre a opinião pública de um determinado serviço ou produto (SALINCA, 2015).

Análise de sentimento é o estudo computacional de opiniões, sentimentos, avaliações, atitudes e emoções expressadas em um texto (JOSE; CHORALIL, 2016). Usualmente se refere a um problema de classificação em que o foco é encontrar e classificar a polaridade das sentenças (ou textos) em negativo, positivo ou neutro (JOSE; CHORALIL, 2016).

Pode-se considerar a análise de sentimento como a utilização de processamento de linguagem natural, mineração de texto e em alguns casos aprendizagem de máquina (JOSE; CHORALIL, 2016). Algumas técnicas de análise de sentimento envolvem o uso de *lexicons* que são dicionários de palavras que

contém pontuações pré-estabelecidas representando o sentimento da palavra (TABOADA et al, 2011).

Existem algumas limitações em relação ao resultado de uma determinada análise de sentimento, pois a acurácia de um algoritmo de análise de sentimento é baseada em quão próxima ela é da análise humana (LIU, 2012).

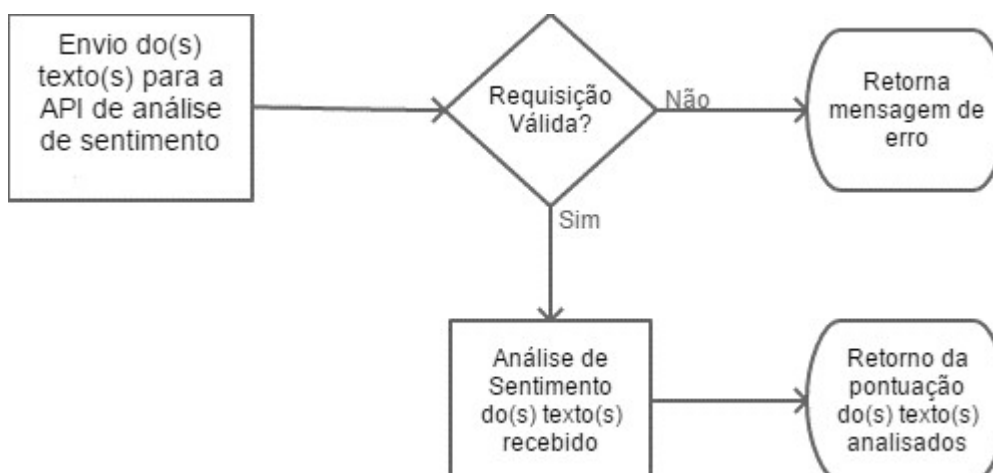
Adicionalmente, segundo Salinca (2015), pode-se observar limitações quando um conjunto de palavras pode ter diferentes significados em contextos variados ou dependendo da linguagem utilizada (Se o algoritmo for baseado na tradução de palavras, a frase pode perder o sentido original) e sentenças com sentimentos ambíguos que normalmente estão presentes em frases sarcásticas.

2.4 FERRAMENTAS DE ANÁLISE DE SENTIMENTO

Na Internet é possível encontrar diversas ferramentas para realizar a análise de sentimento em textos, sentenças ou documentos. Todas as ferramentas utilizadas neste trabalho são pagas, porém possuem versões de teste, que possibilitaram a utilização das mesmas durante a coleta e análise dos dados.

O funcionamento das ferramentas é bastante similar. A Figura 2, apresenta o diagrama que descreve o funcionamento básico das ferramentas (*Microsoft Azure, Repustate e Meaning Cloud*), que consiste basicamente no envio de requisições para um *webservice* que por sua vez retorna uma pontuação da palavra enviada.

Figura 4: Diagrama da Análise de Sentimento



Fonte: Autoria Própria

No restante deste capítulo serão brevemente descritas as diferenças das ferramentas de análise de sentimento, escolhidas com base no fato de serem ferramentas pagas e possuírem utilização semelhante entre si.

2.4.1 Microsoft Azure – *Text Analytics*

A *Microsoft Azure*⁴ é uma plataforma de recursos na nuvem, como banco de dados, servidores, etc. e possui diferentes ferramentas para os mais diversos tipos de problemas e necessidades. A partir dessa variedade, a ferramenta de análise de sentimento foi selecionada.

A ferramenta de análise de sentimento possui uma API que é utilizada para automatizar o processo de classificação de um texto. A API avalia cada texto de entrada fornecido e retorna uma pontuação para cada texto que varia de 0 (muito negativo) até 1 (muito positivo) e pontuações próximas de 0,5 são consideradas neutras (MICROSOFT, 2019). No Quadro 1 é possível observar a relação entre a pontuação e o sentimento extraído do texto.

De acordo com a documentação da API, a partir de uma base grande de dados foi utilizado técnicas de análise de texto e associação de palavras, junto com um algoritmo de classificação, para que fosse possível ao serviço gerar pontuações de positivo e negativo a partir de uma frase, como mencionado no parágrafo anterior.

⁴ <https://azure.microsoft.com/pt-br/>

Quadro 1: Relação da pontuação e sentimento – Microsoft Azure

Pontuação	Sentimento
Pontuação ≤ 0.25	Muito Negativo (<i>Negative+</i>)
Pontuação > 0.25 & < 0.5	Negativo (<i>Negative</i>)
Pontuação ≥ 0.5 & < 0.6	Neutro (<i>Neutral</i>)
Pontuação ≥ 0.6 & < 0.75	Positivo (<i>Positive</i>)
Pontuação ≥ 0.75	Muito Positivo (<i>Positive+</i>)

Fonte: Autoria Própria

Atualmente a API possui suporte aos idiomas: Inglês, Português, Espanhol, Alemão e Francês, enquanto outras línguas estão em fase de desenvolvimento.

2.4.2 Repustate – *Sentiment Analysis*

A API de análise de sentimento da empresa *Repustate*⁵ é compreensível e flexível pelo fato de aceitar novas regras de classificação personalizadas e receber novos dados de treinamento para personalizar a análise de documentos ou textos do usuário. Além disso, a empresa disponibiliza uma documentação clara e objetiva a respeito da ferramenta.

Para realizar a análise de sentimento a API pode executar alguns passos que melhoram a eficiência da análise em alguns casos. De acordo com a documentação, cada passo tem como objetivo obter informações da sentença e são: marcação de parte do texto, lematização, polarização de palavras, negações e amplificadores (como as palavras “muito”, “sempre”, “nunca”, etc).

A marcação de parte do texto busca encontrar construções gramaticais para ajudar o algoritmo de aprendizagem de máquina na classificação. A Lematização identifica palavras repetidas na sentença, mesmo que a mesma palavra possa ser escrita de forma diferente devido a regras do idioma. A polarização de palavras identifica palavras que possuam uma emoção associada. Por fim, são analisadas as negações e amplificadores presentes na sentença.

⁵ <https://www.repustate.com/>

A API avalia cada texto de entrada fornecido e retorna uma pontuação para cada texto que varia de -1 (muito negativo) até 1 (muito positivo) e pontuações próximas de 0 são consideradas neutras. O Quadro 2 mostra a relação entre a pontuação e o sentimento extraído.

Quadro 2: Relação da pontuação e sentimento - Repustate

Pontuação	Sentimento
Pontuação ≤ -0.5	Muito Negativo (<i>Negative+</i>)
Pontuação > -0.5 & < 0	Negativo (<i>Negative</i>)
Pontuação ≥ 0 & < 0.1	Neutro (<i>Neutral</i>)
Pontuação ≥ 0.1 & < 0.5	Positivo (<i>Positive</i>)
Pontuação ≥ 0.5	Muito Positivo (<i>Positive+</i>)

Fonte: Autoria Própria

2.4.3 Meaning Cloud – *Sentiment Analysis*

A API de análise de sentimento da empresa *Meaning Cloud*⁶ também é uma ferramenta versátil pelo fato de aceitar customizações como bases de treino, dicionários de palavras específicas do domínio e ainda modelos de sentimento que sejam mais compatíveis com a necessidade.

O texto fornecido é analisado para determinar se expressa um sentimento positivo, negativo ou neutro. De acordo com a documentação, a polaridade local das diferentes sentenças no texto é identificada e a relação entre elas avaliada, resultando em um valor de polaridade global para todo o texto e a pontuação retornada pela API segue detalhada no Quadro 3.

Quadro 3: Relação da pontuação e sentimento – Meaning Cloud

Pontuação	Sentimento
Pontuação N+	Muito Negativo (<i>Negative+</i>)
Pontuação N	Negativo (<i>Negative</i>)
Pontuação NEU	Neutro (<i>Neutral</i>)
Pontuação P	Positivo (<i>Positive</i>)
Pontuação P+	Muito Positivo (<i>Positive+</i>)

⁶ <https://www.meaningcloud.com/>

Fonte: Autoria Própria

O mais interessante é que, até o presente momento, está em fase de testes a funcionalidade de detecção de ironia. Entretanto, essa fase de testes é fechada, não sendo possível avaliar sua eficácia.

2.5 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo foram apresentados os elementos principais para embasamento teórico e para o desenvolvimento da ferramenta, definindo as ferramentas para realizar a análise de sentimento.

Após analisar as ferramentas, pode-se perceber que possuem um modo de utilização semelhante, porém, cada uma delas utiliza um algoritmo diferente para realizar a análise de sentimento.

3 ESTUDO DE CASO

O estudo de caso foi realizado utilizando mensagens provenientes do Twitter⁷ a fim de servir como base para trabalhos futuros. Observar como as ferramentas produzem resultados diferentes para um mesmo *tweet* é um passo importante para que seja possível definir algum grau de confiabilidade das ferramentas, portanto, nenhuma base de treino ou modelos de classificação externos foram utilizados para que seja feita uma análise dos resultados sem nenhuma especialização das ferramentas.

O resultado esperado deste trabalho é avaliar o desempenho individual das ferramentas utilizadas para realizar a análise de sentimento. Para que em trabalhos futuros seja possível escolher entre uma dessas ferramentas para encontrar o estado sentimental de grupos de interesse no Twitter.

Primeiramente é feito um detalhamento sobre o Twitter e suas funcionalidades para embasar a metodologia do trabalho. Após isso, um detalhamento é feito sobre a ferramenta TScoopws que possibilitou a utilização das três ferramentas de análise de sentimento enquanto eram coletados *tweets*.

3.1 METODOLOGIA

O estudo de caso inicia realizando uma pesquisa descritiva a fim de avaliar o desempenho, características, similaridades e diferenças entre as ferramentas escolhidas para realização da análise de sentimento de *tweets*. Ao fim do trabalho é esperado que ele possa servir como primeiro passo para trabalhos futuros na área.

Inicialmente é feita uma abordagem quantitativa para apontar por meio de números e gráficos as pontuações geradas pelas ferramentas, para que possam ser comparadas e avaliadas. As ferramentas serão avaliadas de acordo com a proximidade que sua avaliação tem com a avaliação humana.

A coleta dos dados se dá pela utilização da API do Twitter para extrair *tweets* de acordo com os parâmetros de busca. Foram realizadas três coletas, nos dias 25 e 26 de março de 2019, 19 e 20 de abril de 2019, 12 e 13 de maio de 2019. Posteriormente, será feito o pré-processamento de cada *tweet* extraído para ser

⁷ <https://twitter.com/>

armazenado no banco de dados e finalmente avaliado por cada ferramenta. A ferramenta TScoopws foi criada para integrar e automatizar todos esses passos citados.

3.2 TScoopws

O TScoopws faz a integração de diferentes ferramentas de análise de sentimento para avaliar o sentimento de *tweets* provenientes de requisições feitas a API do Twitter. A estrutura da ferramenta é dividida em quatro partes: *migrations*, *models*, banco de dados e *controllers*.

O *framework* PHP Laravel⁸ 5 foi escolhido para auxiliar no desenvolvimento da ferramenta pela facilidade de trabalhar com requisições HTTP e integração com o banco de dados Postgres⁹. Além disso, a organização padrão dos pacotes dentro do *framework* auxilia no desenvolvimento permitindo a fácil integração entre as APIs de análise de sentimento.

A ferramenta possui duas partes principais que são utilizadas para cumprir seu objetivo; a busca e pré-processamento são usados para receber *tweets* e preparar os mesmos para serem submetidos às APIs de análise de sentimento e posteriormente a parte de análise de sentimento submete os *tweets* as APIs e armazena os resultados no banco de dados.

3.2.1 Estrutura

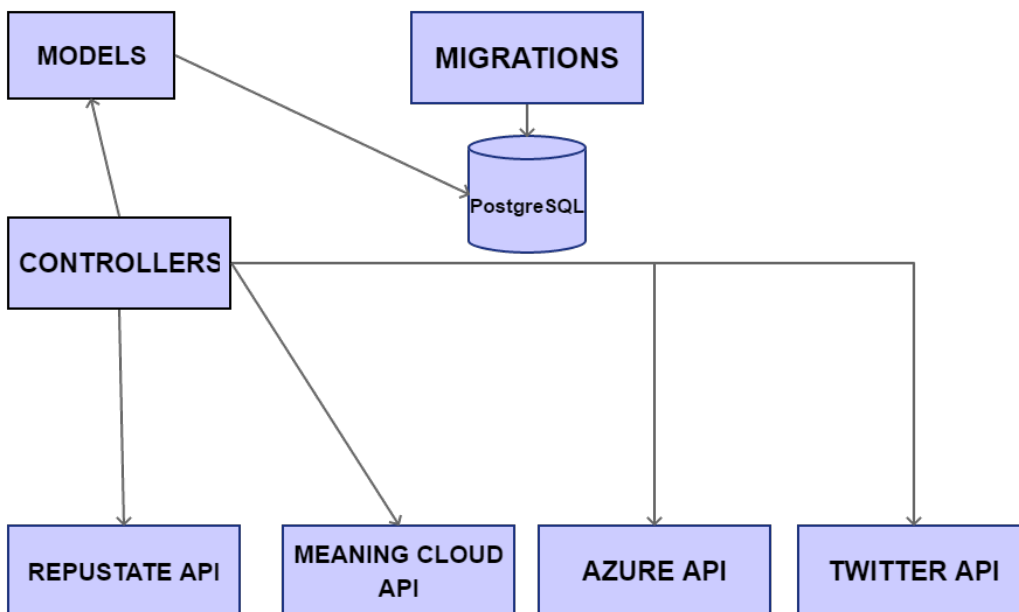
A estrutura do Tscoopws é ilustrada por um esboço na Figura 5, que representa de maneira generalizada como os módulos interagem entre si e com o banco de dados. As *controllers* são responsáveis pela comunicação com todas as APIs e *models*, enquanto *migrations* e *models* se relacionam diretamente com o banco de dados

No restante do capítulo, mais detalhes sobre os módulos serão explorados para melhor entendimento da estrutura da ferramenta.

⁸ <https://laravel.com/>

⁹ <https://www.postgresql.org/>

Figura 5 - Estrutura generalizada da ferramenta



Fonte - Autoria Própria

3.2.1.1 Migrations

Migrations são classes utilizadas para definir regras e estrutura do banco de dados de forma simplificada, ao invés de utilizar o SQL comum para criação das tabelas, o Laravel 5 trabalha com Eloquent ORM (*Object-Relational Mapping*) que é uma técnica para converter dados entre banco de dados relacional e linguagens de programação Orientada a Objetos.

Foi necessária a criação de quatro migrations (cada uma representando uma tabela no banco de dados). Cada migration é executada apenas uma vez antes de o *software* ser iniciado e não são executadas novamente. Na Figura 6 é possível ver a migration utilizada para criação da tabela `tscoop.tweets`. O restante das migrations está junto do código fonte do projeto.

Figura 6 - Migration da tabela tweets

```
class CreateTweetsTable extends Migration {
  public function up() {
    Schema::create('tscoop.tweets', function (Blueprint $table) {
      $table->increments('id');
      $table->string('text', 200);
      $table->dateTime('created_at');
      $table->unsignedInteger('keywords_id');

      $table->foreign('keywords_id')->references('id')->on('tscoop.keywords');
    });
  }

  public function down()
  {
    Schema::dropIfExists('tscoop.tweets');
  }
}
```

Fonte: Autoria Própria

O método `up()` é executado na criação da *migration* e contém a informação necessária em relação aos atributos e relacionamentos da tabela, que posteriormente é criada no banco de dados. Caso seja necessário desfazer as alterações é feita uma chamada para o método `down()` que realiza o *rollback* na base de dados.

3.2.1.2 Models

Depois que as *migrations* são criadas, são necessárias classes para representar as tabelas do banco de dados e realizar a conexão com o banco de dados. Então foram criadas as *models*: *Tweet*, *Sentiment*, *Keyword* e *Tool*. A Figura 7 mostra o código da *Model Tweet*.

Figura 7 - Código da Model Tweet

```

class Tweet extends Model {
    protected $fillable = [
        'text',
        'created_at'
    ];

    public $timestamps = false;

    protected $table = 'tscoop.tweets';

    protected $with = 'keywords';

    public function keywords() {
        return $this->belongsTo(Keyword::class, 'keywords_id', 'id');
    }

    public function sentiment(){
        return $this->hasOne(Sentiment::class, 'tweets_id', 'id');
    }
}

```

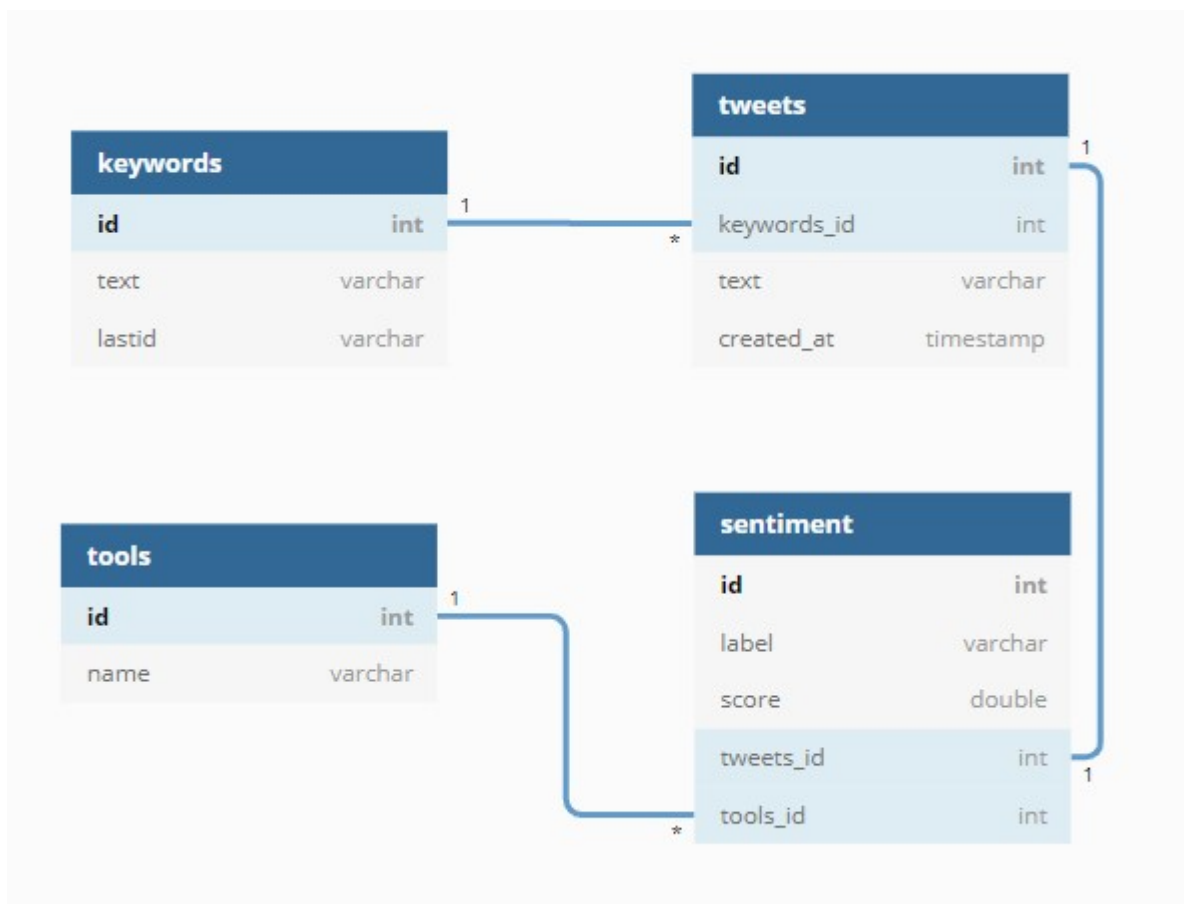
Fonte: Autoria Própria

A classe *Tweet* estende a classe *Model* que contém os métodos de acesso e *drivers* de conexão com o banco de dados. A variável `$fillable` contém os atributos que podem ser passados por requisições para aumentar a segurança e evitar vulnerabilidades de atribuições em massa. No restante da implementação da classe, é atribuído o nome da tabela da *model* no banco de dados e os relacionamentos.

3.2.1.3 PostgreSQL

A Figura 8 mostra um esboço do modelo relacional das tabelas do banco de dados. A tabela *tweets* armazena todos os *tweets* recebidos da API do Twitter e sua data de criação e o atributo *keywords_id* é uma referência para relacionar o *tweet* recebido de acordo o termo de busca (*keyword*). Por sua vez a tabela *keywords* armazena todos os termos de busca utilizados para manter o histórico.

Figura 8 - Modelo relacional do banco de dados



Fonte: Autoria Própria

A tabela *sentiment* armazena a avaliação de sentimento feita pela ferramenta de análise de sentimento, cujos nomes são armazenados na tabela *tools*. A tabela sentimento contém a informação sobre a pontuação de sentimento de um *tweet* e um rótulo associado a ele.

3.2.2 Busca e pré-processamento

A primeira função do *software* ao ser iniciado é realizar uma busca por *Tweets* na API de busca padrão do Twitter utilizando a biblioteca Twitter, desenvolvida pelo usuário ThuJohn¹⁰ disponível no Github. A Figura 9, mostra a *query* criada antes de realizar a requisição para a API do Twitter.

¹⁰ <https://github.com/thujohn/twitter>

Figura 9 – Query de busca na API do Twitter

```
$query = [
    'q' => $searchWord,
    'count' => 100,
    'include_entities' => false,
    'result_type' => 'mixed',
    'lang' => 'pt',
    'until' => Carbon::now()->addDay()->format('Y-m-d')
];
```

Fonte: Autoria Própria

O parâmetro *until* contém sempre a data do dia seguinte para retornar *tweets* criados até a data presente de execução. Após isso, é feita a requisição para a API e retornado um conjunto de *tweets* relacionados aos termos de busca. Este conjunto é então processado de acordo com o código na Figura 10, para ser então armazenado no banco de dados e posteriormente analisado pelas ferramentas de análise de sentimento.

Figura 10 – Laço principal de iteração sobre o conjunto de tweets

```
foreach ($result as $key){
    if(isset($key)){
        foreach ($key as $obj){
            if (isset($obj->text)){
                if (strpos($obj->text, $searchWord) !== false){
                    $tweet = new Tweet();
                    $tweet->text = $this->parseTweet($obj->text);
                    $tweet->created_at = Carbon::parse($obj->created_at);
                    $tweet->created_at = $tweet->created_at->format('d-m-Y H:i:s');
                    $tweet->keywords_id = $keyword['id'];
                    $tweet->save();

                    $this->analyzeSentence($tweet);
                }
            }
        }
    }
}
```

Fonte: Autoria Própria

À primeira vista a existência de laços aninhados pode parecer um pouco complexa, porém isso se faz necessário, pois a API de busca do Twitter retorna vetores dentro de vetores em seus objetos. Após o passo anterior, são filtrados *tweets* que contêm a palavra utilizada na busca a fim de evitar *tweets* não relacionados ao tema pesquisado (vide Figura 11).

Figura 11 - Verificação da existência da palavra de busca e armazenamento do tweet

```

if(strpos($obj->text, $searchWord) !== false){
    $tweet = new Tweet();
    $tweet->text = $this->parseTweet($obj->text);
    $tweet->created_at = Carbon::parse($obj->created_at);
    $tweet->created_at = $tweet->created_at->format('d-m-Y H:i:s');
    $tweet->keywords_id = $keyword['id'];
    $tweet->save();

    $this->analyzeSentence($tweet);
}

```

Fonte: Autoria Própria

Após realizar essa verificação, um objeto da *model Tweet* é criado na variável *\$tweet* e feito o pré-processamento (Figura 12) para remover *links* http(s) e acentos. Então são criados os atributos correspondentes aos existentes no banco de dados e salvos no banco de dados pelo método *save()*, para que posteriormente seja submetido para análise de sentimento por meio das APIs citadas no capítulo 2.4.

Figura 12 – Código do método *parseTweet()*

```

protected function parseTweet($tweet) {
    $replace = '/\b(https?):\/\/[-A-Z0-9+&@#\/%?~_!:\.,;]*[A-Z0-9+&@#\/%?~_!$]/i';

    $tweet = preg_replace($replace, '', $tweet); // This removes the https link that comes with a
    few tweets.
    $tweet =
    preg_replace(array("/(á|à|ã|â|ä)/", "/(Á|À|Ã|Â|Ä)/", "/(é|è|ê|ë)/", "/(É|È|Ê|Ë)/", "/(í|î|ï|ì)/", "/(Í|
    Î|Ï|Ï)/", "/(ó|ò|ô|õ|ö)/", "/(Ó|Ò|Ô|Õ|Ö)/", "/(ú|ù|û|ü)/", "/(Ú|Ù|Û|Ü)/", "/(ñ)/", "/(Ñ)/"),
    explode(" ", "a A e E i I o O u U n N"), $tweet); // This removes any accent in the tweet.

    return $tweet;
}

```

Fonte: Autoria Própria

3.2.3 Análise de sentimento

O método *analyzeSentence()* serve como um facilitador para as chamadas individuais de cada método que realiza uma requisição para uma API de análise de sentimento, a Figura 13 mostra o código do método, o qual também foi criado para facilitar a inclusão de eventuais novas APIs de análise de sentimento.

Figura 13 – Código do método `analyzeSentence`

```
public function analyzeSentence($tweet) {
    $this->analyzeSentenceAzure($tweet);
    $this->analyzeSentenceMeaningCloud($tweet);
    $this->analyzeSentenceRepustate($tweet);
}
```

Fonte: Autoria Própria

O objetivo desse método é concentrar as chamadas as APIs de análise de sentimento, para facilitar a inclusão e remoção de novas APIs. Antes ainda de as requisições serem feitas, é preciso inicializar as classes. Na Figura 14 é possível ver o código do método `analyzeSentenceAzure()`, que é responsável por inicializar a classe Azure e chamar o método que realiza a requisição para a API da *Microsoft Azure*.

Figura 14 – Código do método `analyzeSentenceAzure`

```
protected function analyzeSentenceAzure($tweet) {
    $azure = new Azure('4918c7c7f8354c3b8783b8bc0e4e3a62', $tweet);
    $azure->sentiment();
}
```

Fonte: Autoria Própria

Todas as classes de análise de sentimento têm a inicialização semelhante, pois recebem no primeiro parâmetro uma chave de identificação única, que é gerada no momento do cadastro no *site* da ferramenta, que funciona como uma senha para autenticar o remetente da requisição e o segundo parâmetro é o próprio *tweet* a ser analisado.

Para cada API de análise de sentimento foi criada uma classe, que tem como objetivo enviar o *tweet* para avaliação e armazenar no banco de dados o resultado dessa avaliação. As Figuras 15 e 16 mostram o código das funções `sentiment()` e `storeSentiment()` respectivamente.

Figura 15 – Código do método sentiment()

```
public function sentiment() {
    $context = stream_context_create ($this->options);
    $result = file_get_contents ($this->url, false, $context);
    $result = json_decode($result, true);
    $result = $result['documents'];
    $result = $result[0];
    $score = $result['score'];

    $this->storeSentiment($score);
}
```

Fonte: Aatoria Própria

O método *sentiment()* utiliza o cURL¹¹ para criar requisições HTTP e realiza a requisição para a API de análise de sentimento pelo método *file_get_contents()*. A variável *\$result* armazena o resultado da requisição e logo em seguida faz o tratamento da mesma, para armazenar na variável *\$score* a pontuação recebida pela API e fazer uma chamada ao método *storeSentiment()*.

Figura 16 – Código do método storeSentiment()

```
public function storeSentiment($score) {
    $tool = Tool::where('name', 'Microsoft Azure')->first();
    $sentiment = new Sentiment();
    $sentiment->tools_id = $tool->id;
    $sentiment->score = $score;

    if ($sentiment->score <= 0.25) {
        $sentiment->label = 'Negative+';
    }
    if ($sentiment->score > 0.25 && $sentiment->score < 0.5) {
        $sentiment->label = 'Negative';
    }
    if ($sentiment->score >= 0.5 && $sentiment->score < 0.6) {
        $sentiment->label = 'Neutral';
    }
    if ($sentiment->score >= 0.6 && $sentiment->score < 0.75) {
        $sentiment->label = 'Positive';
    }
    if ($sentiment->score >= 0.75) {
        $sentiment->label = 'Positive+';
    }

    $sentiment->tweets_id = $this->tweet->id;
    $sentiment->save();
}
```

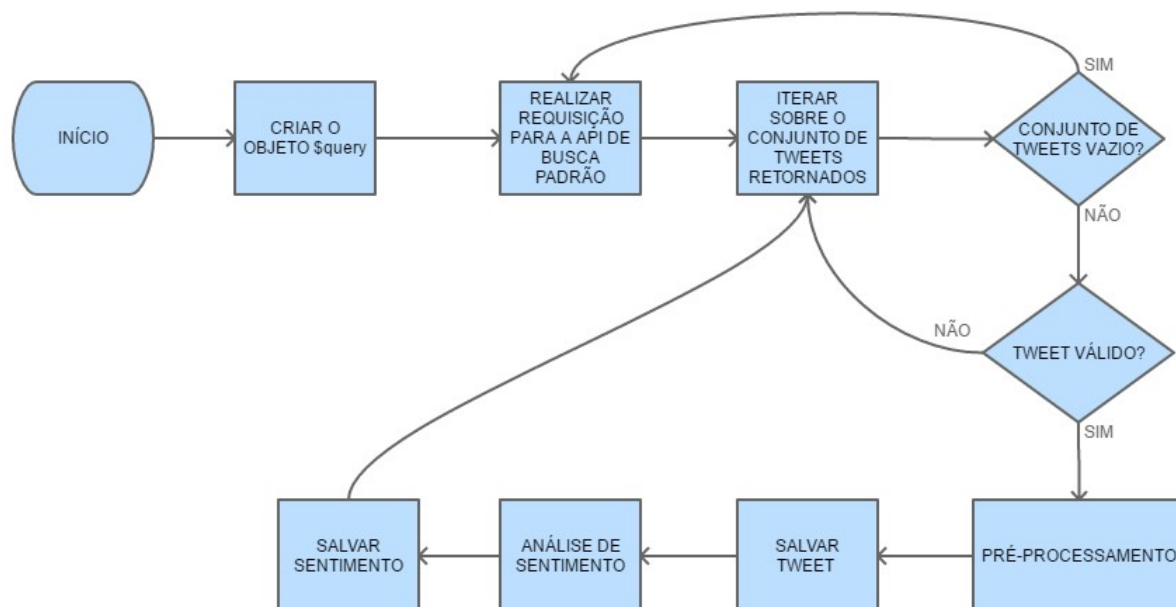
Fonte: Aatoria Própria

¹¹ <https://curl.haxx.se/>

A principal função do método *storeSentiment()* é classificar um *tweet* baseado em sua pontuação de acordo com a Quadro 1 para então armazenar no banco de dados. Essa classificação é importante para uma visão generalizada do estado da rede após descobrir qual a média entre as pontuações.

Esses passos são executados de maneira similar para todos os métodos de análise de sentimento que estiverem presentes no método *analyzeSentence()* (ver Figura 13) e posteriormente o próximo *tweet* e a próxima interação do laço é executada (ver Figura 10) até o fim do conjunto de *tweets* retornado pela API. Para melhor entendimento do fluxo do TScoopws, a Figura 17 contém um diagrama da execução da ferramenta.

Figura 17 – Diagrama de execução do TScoopws



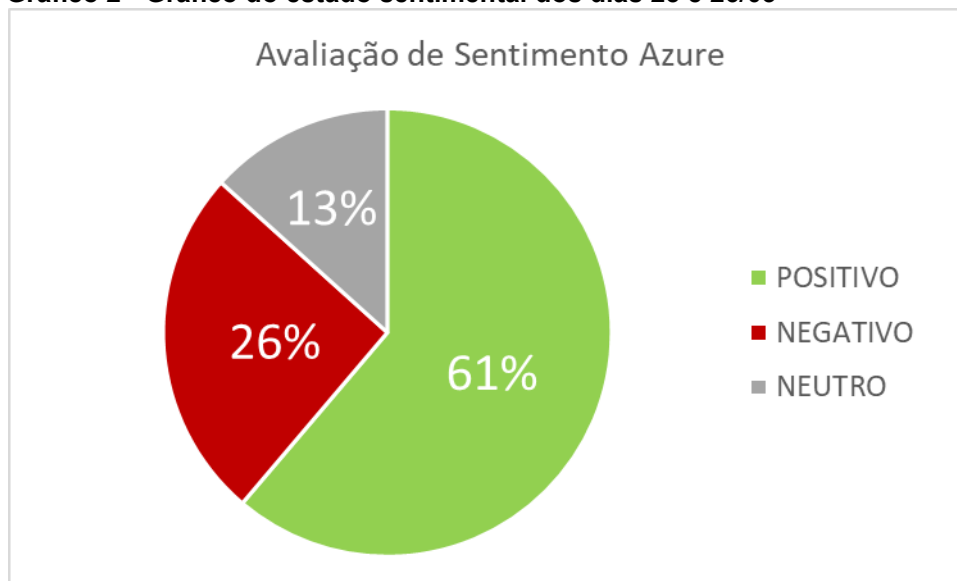
Fonte: Autoria Própria

3.3 RESULTADOS

A primeira coleta do *software* ocorreu entre os dias 25 e 26 de março de 2019, essa data se mostrou promissora devido a prisão do ex-presidente Michel Temer. Portanto, o termo de busca do *software* nesses dias foi a palavra “temer”. Nesse momento da coleta, ainda não havia sido incluído no *software* as APIs *Meaning Cloud* e *Repustate*, portanto só foi considerada a API do Azure e considerado apenas os sentimentos positivo, negativo e neutro.

Como a API de busca padrão do Twitter não provê completude dos dados, foi feita a avaliação de sentimento para essa amostra que contém 3090 *tweets*. O Gráfico 2 mostra as porcentagens de avaliação de sentimento dos *tweets* e podemos ver que o estado sentimental dessa parcela extraída do Twitter é, em sua maior parte, positivo.

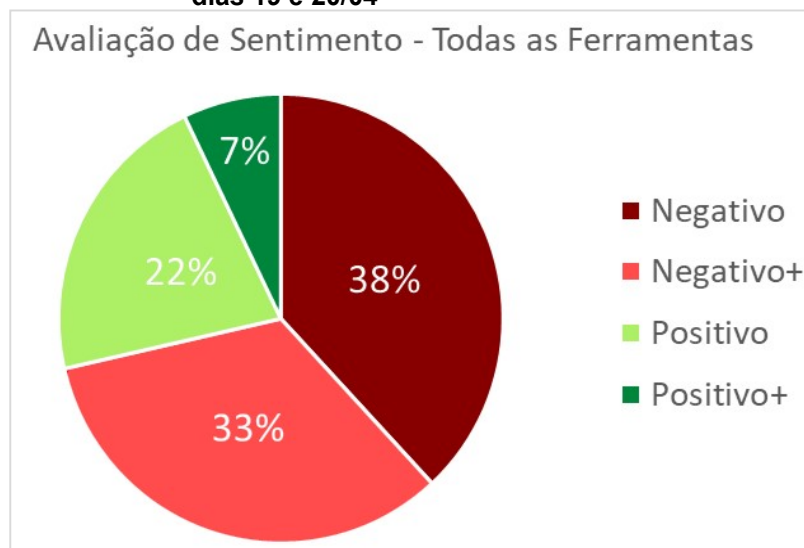
Gráfico 2 - Gráfico do estado sentimental dos dias 25 e 26/03



Fonte: Autoria Própria

Após essa coleta, foi realizada outra coleta utilizando os mesmos parâmetros de busca por *tweets* nos dias 19 e 20 de abril, com objetivo de identificar possíveis mudanças de estado na rede e utilizar duas novas APIs (*Meaning Cloud* e *Repustate*) juntamente com a API do *Microsoft Azure*. O Gráfico 3 mostra a média da avaliação das 3 APIs em relação a uma amostra de 252 *tweets*.

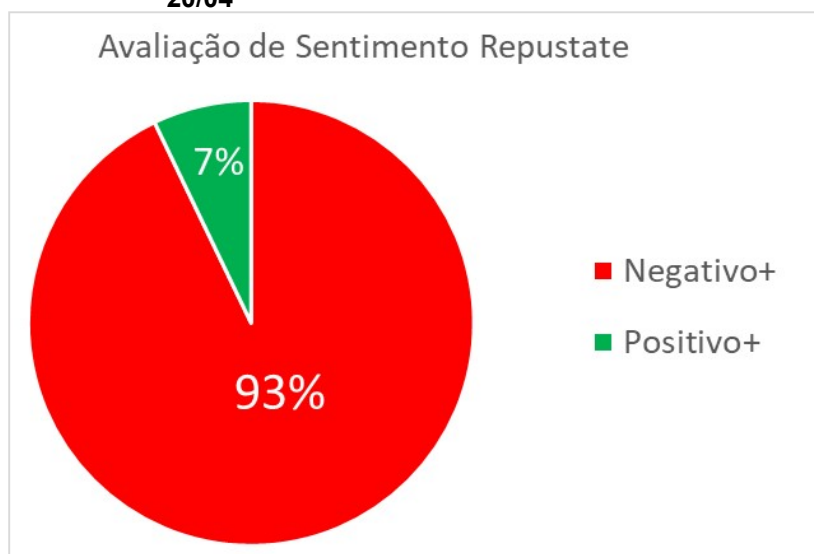
Gráfico 3 – Média da avaliação de sentimento de todas as ferramentas nos dias 19 e 20/04



Fonte: Autoria Própria

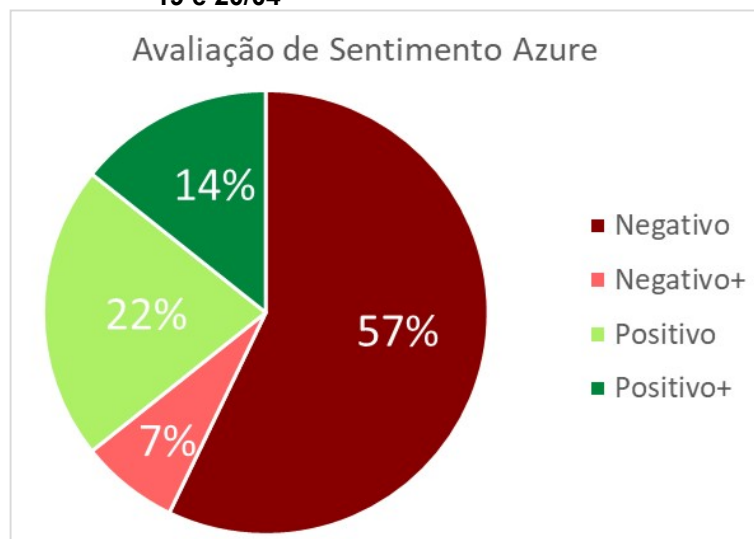
Para se perceber melhor a diferença das diferentes classificações obtidas pelas ferramentas, os Gráficos 4, 5 e 6 foram colocados para apresentar melhor as classificações individuais de cada ferramenta. É importante analisar esses gráficos para perceber como para o mesmo conjunto de *tweets* as ferramentas podem gerar um estado sentimental da rede completamente diferente da outra.

Gráfico 4 - Avaliação de Sentimento da ferramenta Repustate nos dias 19 e 20/04



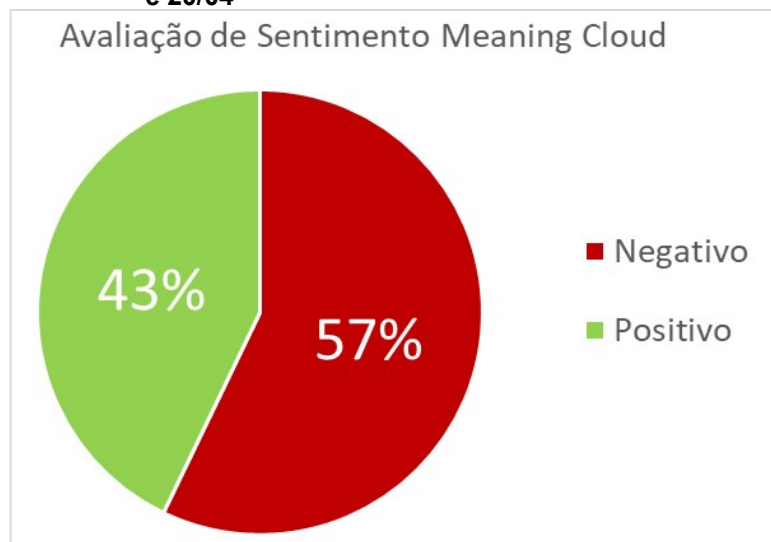
Fonte: Autoria Própria

Gráfico 5 - Avaliação de Sentimento da ferramenta Microsoft Azure nos dias 19 e 20/04



Fonte: Autoria Própria

Gráfico 6 - Avaliação de Sentimento da ferramenta Meaning Cloud nos dias 19 e 20/04



Fonte: Autoria Própria

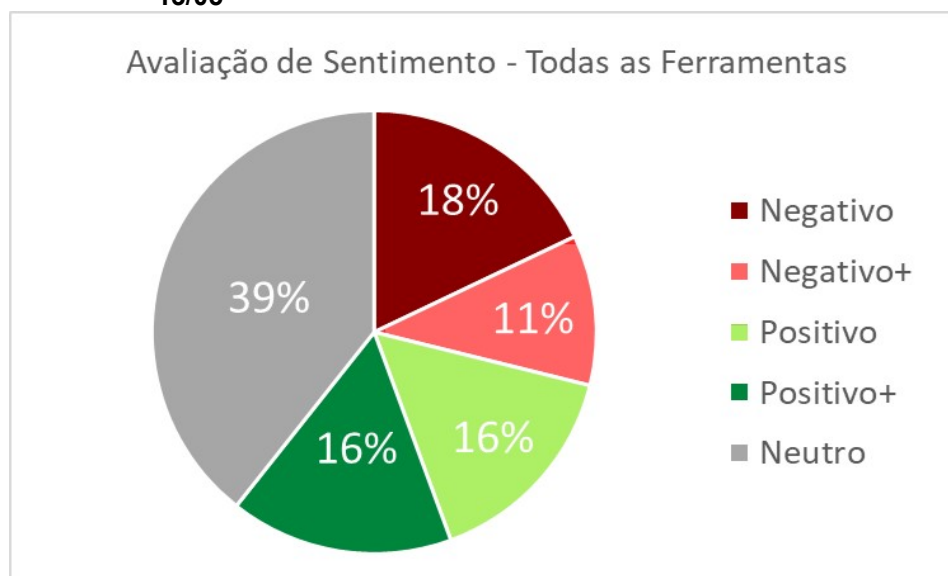
Nesse conjunto de *tweets* submetidos, é possível ver que a ferramenta *Repustate* classificou todos os *tweets* como *Positivo+* ou *Negativo+* e a ferramenta *Meaning Cloud* classificou todos os *tweets* como *Positivo* ou *Negativo*. Descobrir quais os motivos dessa classificação podem gerar melhor entendimento sobre o desempenho da ferramenta para determinadas entradas.

Uma nova coleta foi realizada nos dias 12 e 13 de maio de 2019 quando houveram notícias relacionadas a cortes de verba nas universidades públicas, foi

identificado que seria criada uma quantidade boa de *tweets* para analisar, portanto foi feita uma busca sobre *tweets* relacionados ao presidente Jair Bolsonaro e o termo de busca utilizado foi “bolsonaro” gerando uma amostra de 687 *tweets*.

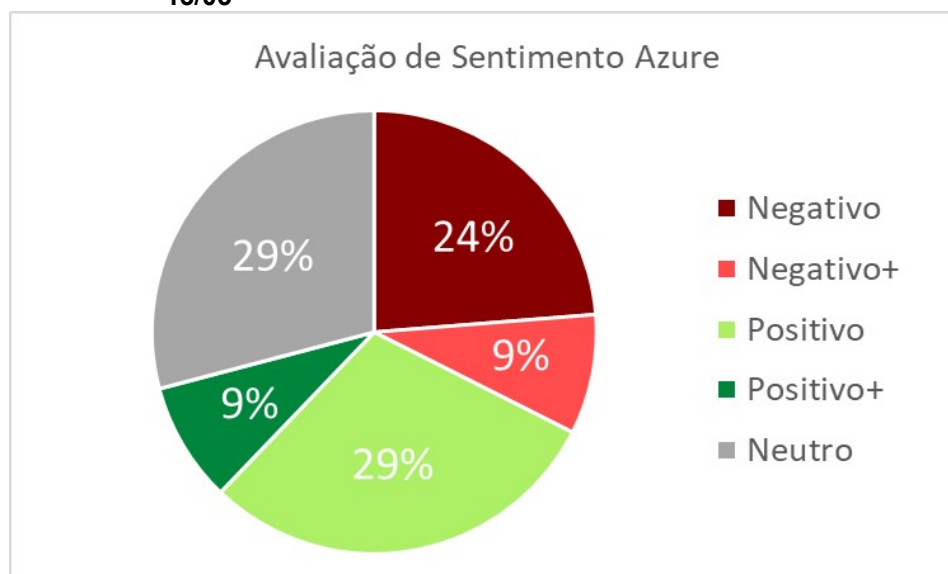
O Gráfico 7 representa a média da avaliação de sentimento de todas as ferramentas, em que é possível notar a polarização bem dividida desse conjunto de *tweets*. Nos Gráficos 8,9 e 10 é possível ver as avaliações individuais de cada ferramenta.

Gráfico 7 - Média da avaliação de sentimento de todas as ferramentas nos dias 12 e 13/05



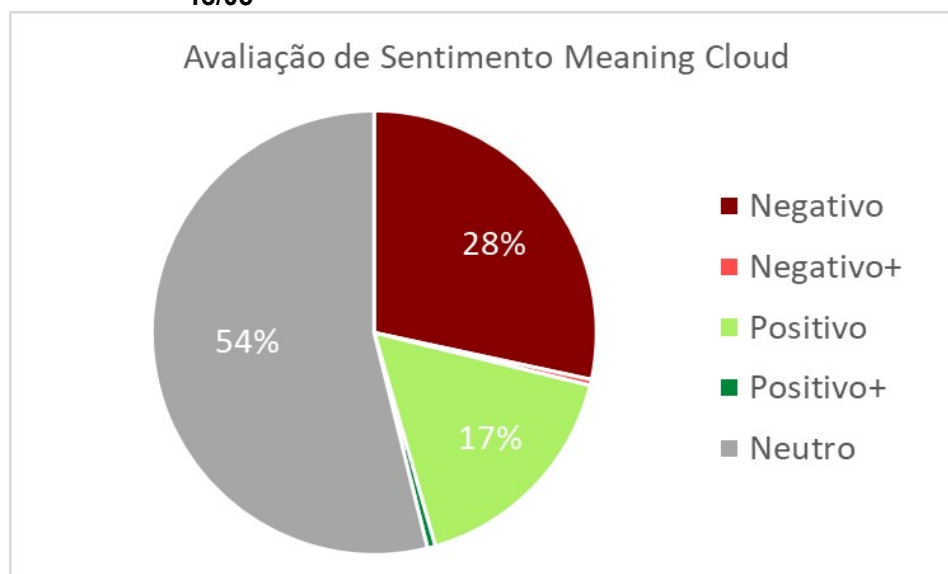
Fonte: Autoria Própria

Gráfico 8 - Avaliação de Sentimento da ferramenta Microsoft Azure nos dias 12 e 13/05



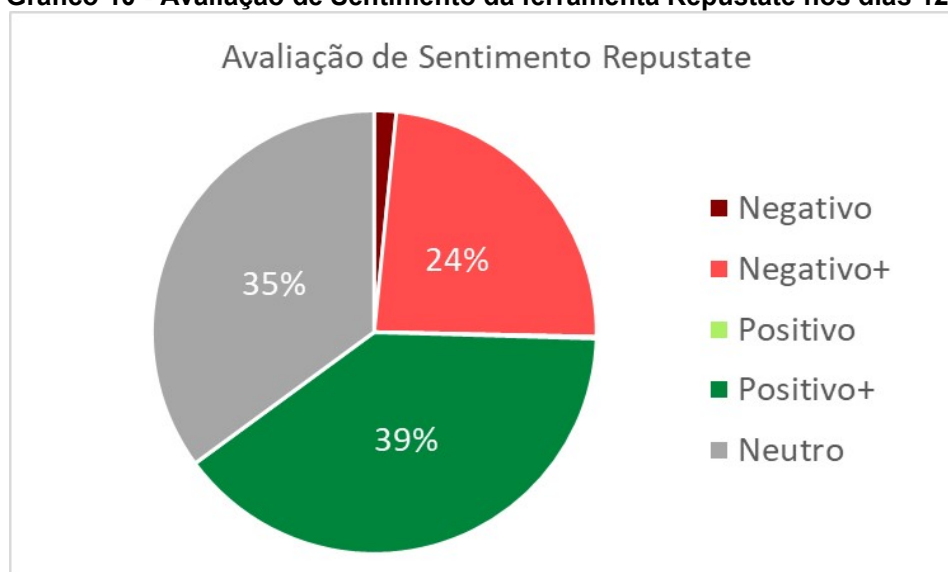
Fonte: Autoria Própria

Gráfico 9 - Avaliação de Sentimento da ferramenta Meaning Cloud nos dias 12 e 13/05



Fonte: Autoria Própria

Gráfico 10 - Avaliação de Sentimento da ferramenta Repustate nos dias 12 e 13/05



Fonte: Autoria Própria

É notável que as ferramentas não avaliam da mesma forma o mesmo conjunto de *tweets* e geram estados sentimentais diferentes devido a implementação individual de cada ferramenta. Para melhor entendimento, os Quadros 4 e 5 demonstram as diferenças na classificação de dois *tweets*.

O *tweet* 1 avaliado no Quadro 4 foi : “@manoKowa159 ai dani pq as musica dele nem e legal e ele apoia o bolsonaro kdkfjk mas vc nao e nazista nao vc e um ANJO te@amo” e o *tweet* 2 avaliado no Quadro 5 foi: “RT @AurelianoGJr: Concordo

com Aecio: @jairbolsonaro e so um principiante... Nao e um ladrao, nao e corrupto, nao recebe propina, nao esta..."

Quadro 4: Avaliação de sentimento das ferramentas em relação ao tweet 1

Ferramenta	Pontuação	Sentimento
Microsoft Azure	0,1947	Negative+
Repustate	0,975	Positive+
Meaning Cloud	0,5	Positive

Fonte: Autoria Própria

Quadro 5: Avaliação de sentimento das ferramentas em relação ao tweet 2

Ferramenta	Pontuação	Sentimento
Microsoft Azure	0.0767	Negative+
Repustate	0,975	Positive+
Meaning Cloud	0	Neutral

Fonte: Autoria Própria

Percebe-se que há uma diferença muito grande da avaliação entre as ferramentas e é possível observar que existem casos em que um *tweet* é avaliado como *Negative+* por uma ferramenta e *Positive+* por outra, sendo dois extremos da classificação.

Adicionalmente, esses resultados mostram que para determinados casos de entrada na ferramenta, ela pode tomar decisões inesperadas ou até mesmo erradas, tornando assim, nenhuma das ferramentas acima completamente confiável para realização de análise de sentimento no Twitter.

4 CONCLUSÃO

A partir dessa diferença na avaliação de sentimento pelas ferramentas, nenhuma delas pode ser completamente confiável devido ao fato de ocorrerem situações em que um mesmo *tweet* é avaliado em duas extremidades de sentimento. Porém a inclusão de mais ferramentas de análise de sentimento pode auxiliar na correta classificação de sentimento desses *tweets* considerando os sentimentos em comum que apresentarem.

A inclusão de novas ferramentas de análise de sentimento deve ser realizada para aumentar a assertividade em relação à classificação, então adicionar ferramentas que realizam classificações muito distintas da avaliação humana pode piorar a assertividade, pois a avaliação humana é a avaliação que mais se aproxima da realidade.

Se a avaliação de sentimento não for precisa, pode tornar a identificação de influência imprecisa, pois não será possível saber qual o estado da rede e como um *tweet* pode estar influenciando um grupo de usuários, o que pode estar levando a uma alteração de sentimento desses usuários.

Por último, a criação deste projeto permite que seja comparado o quão preciso essas 3 ferramentas de análise de sentimento são. Portanto, um trabalho futuro poderia utilizar esta ferramenta para buscar tópicos em geral negativos (como crimes terroristas) onde espera-se que a maioria das pessoas façam comentários contra este tipo de ação e verificar qual das 3 ferramentas mais se aproximam do esperado. Fazer isso com diferentes temas e concluir, em qual tipo de assunto uma ferramenta se saiu melhor que a outra.

4.1 TRABALHOS FUTUROS

As oportunidades de trabalhos futuros são grandes na área de redes de influência, em que algumas perguntas sobre esses grupos ficam em aberto como por exemplo: “Será que cada conjunto de *tweets* retornados por uma busca, pode ser considerado um grupo de interesse?”.

Estudar sobre como são formados grupos de interesse e como encontrá-los no Twitter, pode ajudar na identificação de influência, pois se for provada a

existência de grupos de interesse, por meio dos assuntos comuns aos grupos seria possível relacionar os mesmos.

Identificar as formas de influência e relacionar com aplicações em Teoria dos Grafos como o conjunto dominante mínimo, pode ajudar a identificar a rede como um todo e determinar quais os pontos onde a influência pode ser mais forte ou mais fraca. Outra aplicação seria verificar se o produto lexicográfico pode representar a influência de um grupo ou pessoa sobre outro.

Utilizando análise de sentimento e analisando o histórico sentimental do grupo, é possível analisar tentativas de influência no mundo real? Sendo assim, ao analisar o histórico pode-se verificar sobre o sentimento em relação a algum produto, pessoa etc. Com isso, seria possível analisar os *tweets* que estão mais presentes ou foram retuitados e como isso tem influência sobre o grupo?

Além disso, seria interessante a possibilidade de acesso a todos os *tweets* para implementação de um sistema multi-agentes, que possa analisar diversos grupos de interesse para inferir seu estado sentimental e histórico.

A partir disso seria possível começar a descobrir tentativas de influência identificando o surgimento de assuntos emergentes e como eles se relacionam com a mudança de sentimento. Identificar os criadores de assuntos emergentes, pode se tornar um trabalho para identificação de perfis falsos, por meio de uma classificação de confiabilidade dos assuntos que possam surgir.

REFERÊNCIAS

BBC. **How much to fake a trend on Twitter?:** In one country, about £150. 2018. Disponível em: <<https://bbc.in/2MUZe55>>. Acesso em: 25 mai. 2018.

CARRILHO JUNIOR, João Ribeiro. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. Disponível em:< <https://bit.ly/2HwKxVn>> . Acesso em 18 abr. 2019.

CHANG JIAKANG, Et al. **What is text mining, how does it work and why is it useful?** 2018. Disponível em: < <https://bit.ly/2WWpaSr> >. Acesso em 17 abr. 2019.

DOJ. **Department of Justice of the United States of America, Russia indictment**, 2018. Disponível em:<<https://bit.ly/2CoyuDD>>. Acesso em: 02 abr. 2018.

FACEBOOK. **Facebook State of Connectivity 2015: A Report on Global Internet Access**, 2016. Disponível em:<<https://bit.ly/1T68UJQ>>. Acesso em: 26 mar. 2019.

FACEBOOK. **Facebook Reports Fourth Quarter and Full Year 2017 Results**, 2018. Disponível em:<<https://bit.ly/2DVuq3l>>. Acesso em: 26 mar. 2019.

FELDMAN, R., Sanger, J. **Text Mining Handbook**. Cambridge, Inglaterra: Cambridge University Press. 2006.

FLURRY. **U.S. Consumers Time-Spent on Mobile Crosses 5 Hours a Day**, 2017. Disponível em:<<https://bit.ly/2lLde2u>>. Acesso em: 28 fev. 2019.

FORBES. **Facebook Says Data On 87 Million People May Have Been Shared in Cambridge Analytica Leak**, 2018. Disponível em:<<https://goo.gl/2qJnce>>. Acesso em: 13 mai. 2018.

LIU, Bing, **Sentiment Analysis and Opinion Mining**. Morgan and Claypool Publishers, May 2012.

SALINCA, Andreea, "**Business reviews classification using sentiment analysis**", 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), pp. 247-250, 2015.

TABOADA, Maite & Brooke, Julian & Tofiloski, Milan & D. Voll, Kimberly & Stede, Manfred. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics. 37. 267-307. 10.1162/COLI_a_00049.

THE GUARDIAN. **Ten key takeaways from Robert Mueller's Russia indictment**, 2018. Disponível em: <<https://bit.ly/2tBp3y2>>. Acesso em: 20 mai. 2019.

THE GUARDIAN. **Should stealth marketing be regulated?** . 2008. Disponível em: <<https://bit.ly/2MX9tpl>>. Acesso em: 14 fev. 2019.

TWITTER. **Q1 2018 Letter to shareholders** . 2018. Disponível em: <<https://bit.ly/2MUfPpu> >. Acesso em: 15 fev. 2019.