

Desenvolvimento de uma Ferramenta para Análise de Sentimentos de Textos Publicados no Twitter

Thayná O. Conceição, Rafael G. Rossi

¹Universidade Federal do Mato Grosso do Sul (UFMS)

UNID. II: Av. Ranulpho Marques Leal, 3484 – CEP 79620-080 – Cx Postal nº 210
Mato Grosso do Sul – MS – Brazil

oliveira.thayna@outlook.com, rafael.g.rossi@ufms.br

Resumo. Devido a popularização da internet e ao surgimento das redes sociais, todos os dias são produzidos milhares de dados em forma de textos, especialmente em ambientes que proporcionam o rápido compartilhamento de mensagens entre usuários, assim como observado no Twitter. Visto que muitos destes textos são opinativos e expressam os sentimentos dos usuários, analisar estas opiniões tem despertado interesse no meio governamental e corporativo, seja para avaliar a reputação de um produto no mercado, como para monitorar a aceitação de determinada campanha política ou popularidade de um candidato. No entanto, realizar esta análise manualmente pode ser um processo bastante custoso e demorado, principalmente quando há uma grande quantidade de textos. Neste cenário, tornam-se necessárias ferramentas baseadas em Mineração de Texto e Análise de Sentimento para organizar, analisar e extrair conhecimento útil dessa massa de dados de maneira mais rápida. Devido a crescente demanda por ferramentas para esse fim e ao fato de a maioria das ferramentas existentes serem limitadas a um domínio específico ou idioma, neste trabalho de conclusão de curso foi proposta uma ferramenta online, denominada Emotion-Monitor, independente de domínio para coletar tweets de interesse e extrair a polaridade dos textos, bem como estatísticas como termos e hashtags mais frequentes, além de um ambiente de interação com o usuário, para que o mesmo consiga refinar os resultados apresentados, caso julgue que a classificação automática ocorreu incorretamente.

1. Introdução

A opinião dos demais indivíduos da sociedade sempre foi item muito relevante durante o processo de tomada de decisão para a maioria das pessoas [Pang et al. 2008]. Muito antes da popularização da internet, era comum pedir recomendações de amigos ou especialistas antes de comprar algum produto ou contratar um serviço, a fim de identificar os possíveis riscos antes de despesar recursos. Qual carro comprar, qual candidato votar na próxima eleição, bem como qual filme assistir no cinema são decisões que comumente tomamos a partir de opiniões e avaliações de terceiros.

Com o surgimento dos blogs, das redes sociais e do crescente compartilhamento de informações *online*, esta prática tem se tornando ainda mais frequente, pois é através das mídias sociais que os usuários acabam compartilhando e expondo opiniões com outros internautas [MacLennan et al. 2014]. Segundo [Pang et al. 2008], cerca de 81% dos

usuários já realizaram alguma pesquisa na internet antes de efetuar a compra de um produto ou serviço e 87% afirmam que a avaliação dos demais usuários influenciou significativamente na tomada de decisão.

Tendo em vista que a internet é um grande veículo para disseminação de textos opinativos, analisá-los e compreender a opinião das pessoas pode ser útil em vários tipos de aplicações, tanto para a área empresarial quanto para a área governamental [Pang et al. 2008], seja para uma empresa mensurar se o novo produto lançado está sendo aceito no mercado, bem como se uma determinada campanha política está sendo bem recebida pelos eleitores. Dessa forma, ambientes que proporcionam o compartilhamento rápido de textos opinativos tem se tornado muito populares, especialmente no âmbito das redes sociais, que concentram uma grande quantidade de usuários trocando informações pessoais, opiniões e avaliações. Esse cenário pode ser observado no *Twitter*¹, pois devido ao seu formato de *microblogging*, que permite que os usuários compartilhem informações com um limite máximo de 140 caracteres², é um ambiente atrativo para pessoas expressarem opiniões rapidamente e, com isso, produzir uma grande quantidade de dados diariamente.

No entanto, quando há uma grande quantidade de dados, realizar esta análise manualmente e extrair conhecimento é um processo muito custoso e demorado. Neste cenário, ferramentas automatizadas para realizar tais tarefas tornam-se necessárias, podendo ser utilizadas técnicas das áreas de Mineração de Textos (MT), que visam reconhecer padrões e extraír informações úteis de documentos não estruturados [Tan et al. 1999] e Análise de Sentimento (AS), que busca criar um modelo computacional capaz de reconhecer emoções a partir de textos [Wilson et al. 2005].

Para realizar esta análise é possível utilizar dois métodos: (i) aprendizado não supervisionado, onde não há a necessidade de utilizar nenhum rótulo previamente definido e não há intervenção do usuário no processo de análise de sentimento; e (ii) aprendizado semissupervisionado, que faz uso de uma base de treinamento previamente rotulada para classificar novos objetos. Ferramentas baseadas em aprendizado não supervisionado costumam utilizar dicionários léxicos com algumas palavras e suas polaridades, classificando novas ocorrências a partir da polaridade individual de cada palavra do texto, enquanto ferramentas semissupervisionadas fazem uso de exemplos rotulados manualmente para induzir a classe do novo objeto.

Dada a crescente demanda por ferramentas capazes de analisar textos oriundos de redes sociais e extraír os sentimentos expressos pelos usuários, surgiram diversas ferramentas para suprir esta necessidade. No entanto, a maioria das ferramentas disponíveis para este fim são limitadas a um domínio específico ou comportam apenas dados no idioma inglês. Além disso, algumas delas não são de uso gratuito e costumam limitar a quantidade de consultas que o usuário pode realizar por mês.

Desta forma, a motivação deste trabalho de conclusão de curso dá-se devido a crescente demanda por ferramentas capazes de processar textos e detectar opiniões, tanto

¹*Twitter*: www.twitter.com/

²Durante a confecção deste artigo, entrou em fase de teste para alguns idiomas a possibilidade de publicar tweets com até 280 caracteres. O limite original de 140 caracteres foi mantido para os idiomas inglês e português, os quais são de interesse desta pesquisa.

no meio acadêmico, político e corporativo. Sendo assim, o objetivo deste trabalho é demonstrar uma visão sobre as diversas estratégias existentes na literatura para análise de sentimento e como aplicá-las em um contexto de redes sociais. Além disso, o presente trabalho tem como resultado o desenvolvimento de uma ferramenta gratuita, independente de domínio e que comporta os idiomas inglês e português, capaz de processar textos extraídos do *Twitter* e apresentar um relatório com as opiniões extraídas, além de estatísticas dos termos e *hashtags* mais frequentes, facilitando assim a compreensão do usuário a cerca dos *tweets* coletados.

O restante deste trabalho está dividido em 6 seções. Na Seção 2 serão abordados alguns conceitos utilizados ao decorrer do trabalho, bem como uma visão geral das técnicas mais utilizadas para a análise de sentimentos. Na Seção 3 serão apresentados alguns trabalhos relacionados. Na Seção 4 serão apresentadas as etapas para a elaboração do projeto, desde o desenvolvimento até as estratégias utilizadas. Na Seção 5 serão apresentados alguns estudos de caso. Por fim, na Seção 6 são apresentadas as considerações finais, quais as contribuições deste projeto e sugestões de trabalhos futuros.

2. Fundamentação Teórica

Nesta seção são demonstrados os conceitos abordados neste trabalho de conclusão de curso e a revisão bibliográfica necessária para o entendimento e desenvolvimento da ferramenta proposta.

2.1. Mineração de Textos

A Mineração de Textos refere-se ao processo de extração de padrões ou conhecimento útil a partir de bases de dados textuais [Aggarwal and Zhai 2012]. Tendo em vista que aproximadamente 80% das informações corporativas estão contidas em documentos textuais [Ur-Rahman and Harding 2012] e 80% do conteúdo *online* também está em formato de texto [Chen 2001], a mineração de textos surge da necessidade de organizar e extrair informações dessa massa textual produzida diariamente.

Na literatura, o processo de mineração pode ser dividido em várias etapas. Neste projeto, no entanto, o processo foi dividido conforme apresentado na Figura 1, baseado no modelo proposto por [Aranha and Passos 2006].



Figura 1. Processo de Mineração de Textos.

Na etapa de coleta, são construídos *crawlers* para consultar e extrair dados das fontes de interesse. No pré-processamento, são utilizadas técnicas para estruturar e padronizar os documentos textuais. Na etapa de indexação, os documentos são identificados

e indexados para possibilitar consultas posteriores. Na etapa de mineração, são aplicados algoritmos para identificar padrões e extrair conhecimento. Por fim, na etapa de análise, os resultados são validados e os resultados apresentados, se houver a necessidade.

No presente trabalho, as etapas do processo de análise de sentimentos a partir de textos publicados no *Twitter* foram desenvolvidas de forma semelhante a estrutura do processo de Mineração de Textos.

2.2. Análise de Sentimentos

A Análise de Sentimentos, também conhecida como Mineração de Opinião, é uma área em grande expansão [Liu 2010] que visa encontrar opiniões e sentimentos contidos em textos a respeito de alguma entidade de interesse [Pang et al. 2008], definindo técnicas automáticas para extrair e apresentar um conhecimento estruturado que possa ser utilizado por um sistema de apoio a tomada de decisão.

Assim como a Mineração de Textos, o processo de classificação utilizado na Análise de Sentimentos também pode ser dividido em diversas etapas. Neste projeto, no entanto, o processo foi dividido em 4 etapas, conforme apresentado na Figura 2, baseado no modelo proposto em [Rodrigues et al. 2010].



Figura 2. Processo de Análise de Sentimentos.

Assim como na Mineração de Textos, na etapa de coleta são feitas as buscas para selecionar os dados que serão analisados no processo, bem como na etapa de pré-processamento estes dados são padronizados e estruturados. Na etapa de classificação, são encontradas as polaridades da opinião, determinando se determinado texto pertence a classe positiva, negativa ou neutra. Na etapa de sumarização, os resultados obtidos são apresentados ao usuário através de gráficos e outras estatísticas, conforme a necessidade da aplicação.

2.3. Aprendizado de Máquina

Aprendizado de Máquina é uma área da Inteligência Artificial que visa possibilitar que o computador reconheça e aprenda padrões existentes de um determinado conjunto de dados [Mitchell 1997], aprendendo com experiências acumuladas de soluções anteriores para induzir a classificação em documentos não rotulados. Dentre os principais métodos utilizados no Aprendizado de Máquina, pode-se destacar:

- **Aprendizado Supervisionado:** este método utiliza apenas exemplos já rotulados, aos quais o usuário já definiu rótulos previamente, para induzir um modelo de classificação para classificar novos objetos [Rossi 2016].

- **Aprendizado Semissupervisionado:** este método realiza a classificação automática de textos dado um conjunto de treino previamente rotulado e um conjunto não rotulado [Rossi 2016].
- **Aprendizado não Supervisionado:** este método não utiliza nenhum rótulo previamente definido para os textos para induzir a classificação.

Neste projeto, foram utilizados os métodos de aprendizado não supervisionado e semissupervisionado para classificar os textos.

3. Trabalhos Relacionados

Nos últimos anos, diversas ferramentas foram construídas para explorar o crescente interesse em analisar opiniões embutidas em textos extraídos do *Twitter*. Nesta seção são apresentadas as ferramentas consideradas mais relevantes e relacionados a este projeto de pesquisa, isto é, ferramentas disponíveis para se realizar a Análise de Sentimentos a partir de textos coletados em redes sociais.

As ferramentas apresentadas são de uso gratuito, assim como a ferramenta desenvolvida nesse projeto. No entanto, além das ferramentas apresentadas nesta seção, existem também outras ferramentas comerciais que seguem a mesma proposta de análise de sentimentos e extração de conhecimento em textos oriundos de redes sociais, como o Scup [Sprinklr 2009] e V-tracker [Viragine et al. 2010]. No entanto, devido ao fato de serem pagas, não foi possível testar suas funcionalidades.

No trabalho de [Silva 2016], foi desenvolvida a ferramenta online Análise de Sentimento, a qual é independente de domínio e analisa sentimentos de textos contidos em redes sociais e alguns portais de notícia. Além de classificar os comentários dos usuários entre comentários positivos, negativos e neutros, é possível visualizar quais sentimentos estão atrelados ao texto, como amor, remorso e mais 20 outros sentimentos, conforme a ontologia apresentada em [Ortony et al. 1990]. Na Figura 3, pode-se observar os resultados apresentados pela ferramenta.



Você pesquisou por: **apple**

Análise atual em relação
ao histórico.
Apresentados em 8

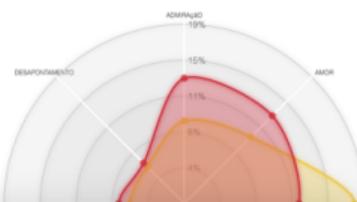


Figura 3. Processo executado para analisar sentimentos a partir de tweets.

Em [Rios et al. 2017], foi desenvolvido o portal TSVIZ para consultar sentimentos a partir de textos extraídos do Twitter, cujo domínio está relacionado principalmente em questões sociais e políticas. Na ferramenta é possível visualizar as palavras mais relevantes relacionadas ao termo consultado, geolocalização das mensagens e a evolução do termo ao longo do tempo através de uma série temporal. Além disso, é possível monitorar qual é a tendência da série temporal para controlar quais são os períodos em que há algum tipo de alteração comportamental dos dados, tendo em vista que estas alterações são comumente associadas a eventos de impacto social e econômico. Na ferramenta também é possível encontrar um gráfico com associação de palavras encontradas no texto. No entanto, as consultas realizadas na aplicação são limitadas há alguns termos de busca previamente definidos. Na Figura 4, pode-se observar os resultados apresentados pela ferramenta TSVIZ.

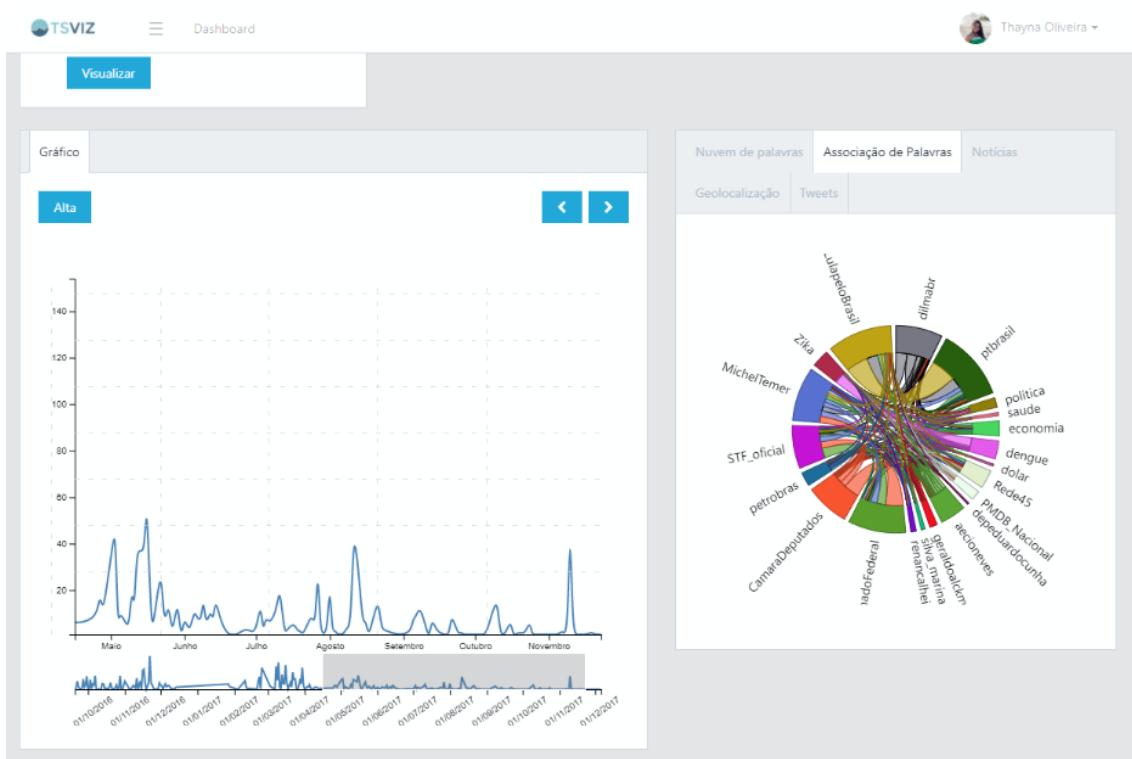


Figura 4. Exemplo de resultados exibidos pela ferramenta TSVIZ.

Em [Araújo et al. 2014], foi desenvolvida uma ferramenta *online*, conhecida como *iFeel*, capaz de analisar sentimentos contidos em textos publicados em redes sociais. A ferramenta utiliza oito métodos de classificação conhecidos na literatura baseados em léxico para realizar a classificação dos textos. Ao final da classificação, é apresentado ao usuário um comparativo dos resultados obtidos em cada método. Dentre os métodos utilizados, No entanto, a ferramenta não apresenta nenhuma informação adicional, limitando os resultados no comparativo de algoritmos. Na Figura 5, pode-se observar os resultados apresentados pela ferramenta.

The screenshot shows the iFeel web application. On the left, there's a dark sidebar with links for Home, Analyse File Texts, and Publications. The main area has a header "iFeel - A Sentiment Analysis Framework" and a message "You should login to analyse multiple sentences". It features a "Login With Facebook" button and "Login" and "Register" buttons. Below this is a search bar with "Language: English" and a dropdown, and a text input field "Type a test" with a "Analyse!" button. The main content area is titled "Methods Results" and shows a table of analysis methods for the input "gremio". The table has columns for Method Name, Status, Method Score, and Polarity.

Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	0	Neutral
SENTISTRENGTH	Completed	0	Neutral
SOCAL	Completed	0	Neutral
HAPPINESSINDEX	Completed	0	Neutral
SANN	Completed	0	Neutral
EMOTICONSDS	Completed	-1	Negative
SENTIMENT140	Completed	0	Neutral
STANFORD	Completed	0	Neutral

Figura 5. Exemplo de resultados exibidos pela ferramenta iFeel.

Dentre as diversas ferramentas disponíveis, temos também a ferramenta desenvolvida em [Jonn 2008], que é uma plataforma gratuita e que tem por finalidade realizar uma busca em blogs, notícias e redes sociais, resultando em indicadores como termos frequentes e a quantidade de postagens por minuto. Nestas análises, também é possível encontrar um gráfico de análise de sentimento que relaciona a palavra procurada com opiniões positivas, negativas e neutras. No entanto, a busca considera apenas termos da língua inglesa. Na Figura 6, pode-se observar os resultados apresentados pela ferramenta SocialMention.

The screenshot shows the SocialMention search interface. At the top, there are navigation links for Blogs, Microblogs, Bookmarks, Images, Video, All, and a search bar containing "bitcoins" with a "Search" button and links for Advanced Search and Preferences. On the left, there's a sidebar with metrics: 72% strength, 1:1 sentiment, 37% passion, 45% reach, 28 seconds avg. per mention, last mention 1 minute ago, 111 unique authors, and 27 retweets. Below this is a "Sentiment" section with a horizontal bar chart showing positive (23), neutral (141), and negative (21) counts. The main content area is titled "Mentions about bitcoins" and displays a list of 15 mentions out of 185. Each mention includes a timestamp, the user who posted it, and a link to the original tweet or post. The first few mentions are:

- RT @imnotgerald: Selling bitcoins for \$10 each. Limited stock. HMU ASAP <https://t.co/6FZkeGX0sr>
twitter.com/pritadewii/status/943149525740302337
1 minute ago - by @pritadewii on [twitter](#)
- RT @AxelMerk: Not quite. Bitcoins may be limited, but there is no limits on either forks or competing crypto currencies <https://t.co/U4wnT8...>
twitter.com/feedsh/status/943149515866906624
1 minute ago - by @feedsh on [twitter](#)
- Sais tu pourquoi je ne crois pas à la réussite du Bitcoins Louis? C'est simple, c'est parce que Jeff Fillion y croi... <https://t.co/ZVVAPGcvpV>
twitter.com/pilote_alain/status/943149506316718081
1 minute ago - by pilote_alain on [twitter](#)

Figura 6. Exemplo de resultados exibidos pela ferramenta SocialMention.

Com base nos trabalhos apresentados, é possível identificar que a maioria das pesquisas conduzidas nesta área buscam extrair conhecimento a partir de um domínio definido, isto é, com termos de busca definidos ou são limitados a um único idioma. Este projeto, no entanto, visa oferecer uma ferramenta independente de domínio, isto é, em vários cenários, e pode ser aplicada para os idiomas inglês e português, implantados neste projeto, ou para qualquer outro idioma que possua um dicionário léxico.

4. Projeto de Implementação

No presente trabalho, o processo de análise de sentimentos implantado na ferramenta proposta, denominada EmotionMonitor, foi dividido em cinco etapas:

1. *Input* do Usuário;
2. Seleção de Dados;
3. Pré-processamento;
4. Análise de Sentimentos;
5. Apresentação do Conhecimento.

Cada etapa desempenha um papel específico para o resultado da análise, de forma que: (i) na etapa de *input* do usuário é definido o termo de busca desejado para coletar os tweets; (ii) na seleção de dados é necessário coletar a base de dados que será utilizada para a análise, a partir da *string* de busca informada pelo usuário; (iii) no pré-processamento estes dados são organizados e padronizados, a fim de tornar a classificação mais promissora; (iv) na etapa de análise de sentimentos os dados já estruturados passam por um processo de detecção de sentimentos; e (v) após classificados, os dados são organizados e apresentados ao usuário na etapa de apresentação do conhecimento. Na Figura 7 é possível visualizar as etapas e o fluxo de informação entre as etapas definidas nesse trabalho de conclusão de curso. Nas próximas seções, todas as etapas são apresentadas com maior detalhamento.

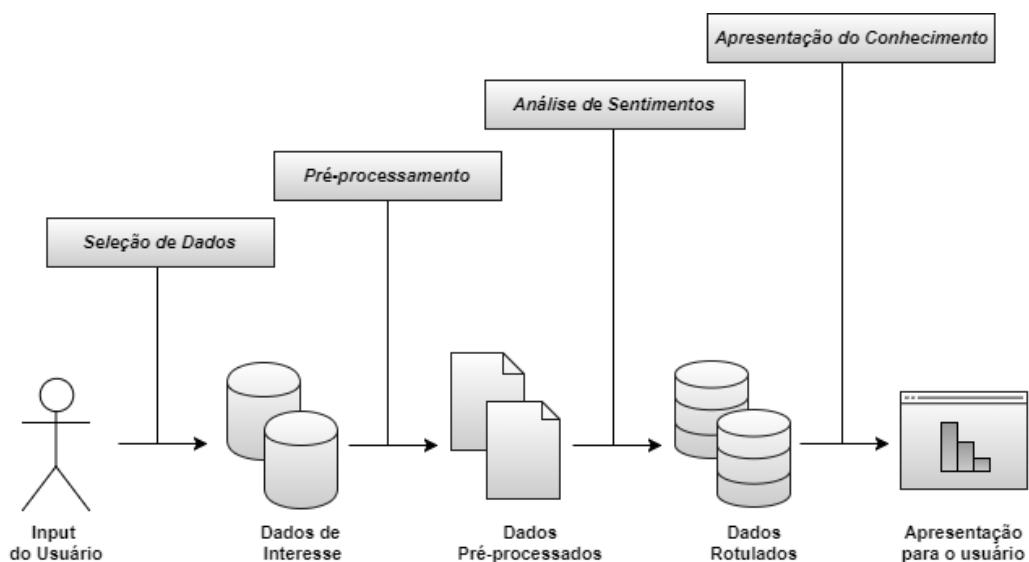


Figura 7. Processo executado para analisar sentimentos a partir de tweets.

4.1. Input do Usuário

A ferramenta desenvolvida é independente de domínio, de forma que é possível realizar consultas a partir de qualquer termo de interesse e ela será capaz de classificar a opinião dos usuários a respeito do tema desejado.

Na tela inicial da aplicação, existe um campo onde o usuário pode informar o termo de busca desejado e qual é o idioma da consulta, pois a ferramenta foi limitada para realizar consultas apenas em *tweets* cujo idioma seja português ou inglês, devido a existência de dicionários léxicos nesses idiomas e de técnicas de pré-processamento de textos para tais linguagens, as quais serão utilizadas para realizar a análise de sentimentos utilizando aprendizado de máquina e para apresentação dos resultados, delimitando a abrangência da aplicação. Na Figura 8 é possível visualizar a *interface* disponibilizada ao usuário.

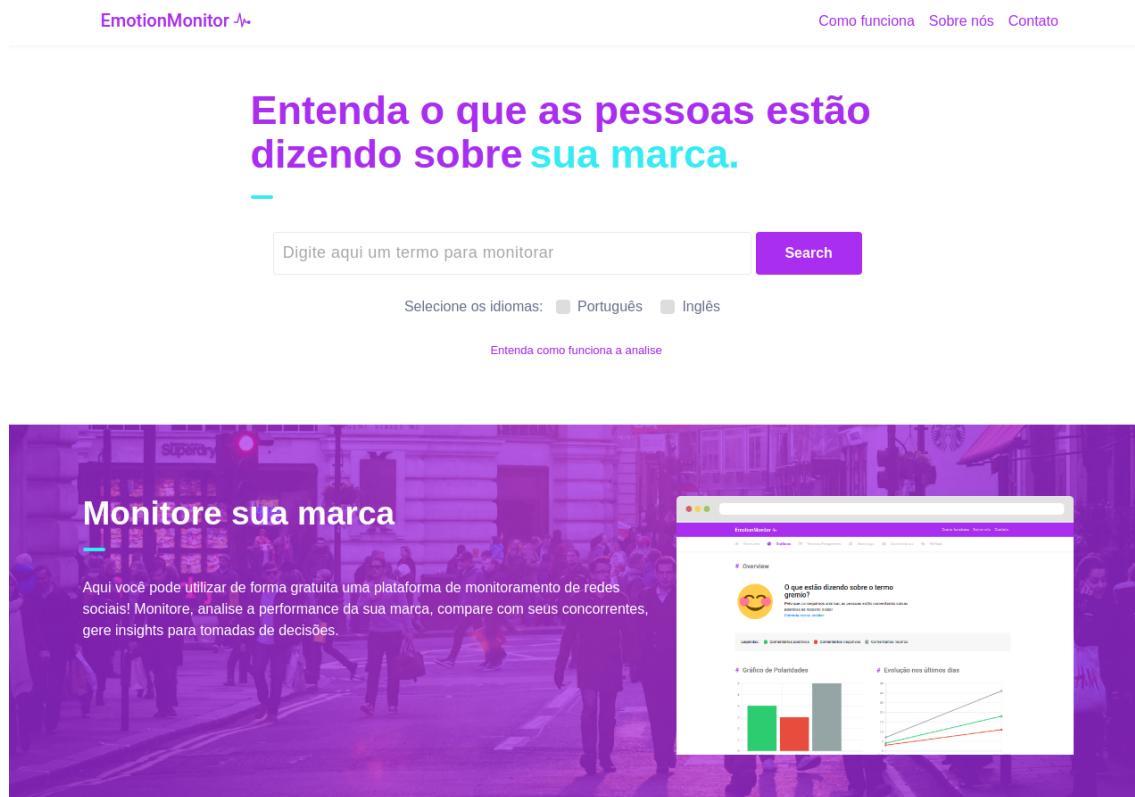


Figura 8. Interface para coleta do termo de busca.

4.2. Seleção de Dados

Após o usuário informar o termo de busca desejado, é necessário construir uma base de dados com *tweets* relacionados ao domínio escolhido para realizar a Análise de Sentimentos. O *Twitter* foi utilizado como fonte de dados para este projeto em virtude da facilidade para consultar *tweets* e da grande quantidade de ferramentas disponíveis para este fim. Diferente de outras redes sociais, no *Twitter* as postagens dos usuários tem por padrão o compartilhamento público e são visíveis para todos, tornando assim muito mais acessível a extração destes dados. Além disso, há uma grande quantidade de ferramentas desenvolvidas que permitem a manipulação e consulta dos *tweets* da rede social.

Neste projeto, para coletar os documentos desejados, foi desenvolvida uma API em JAVA, denominada TwitterAPI, que recebe como parâmetro o termo de busca informado pelo usuário, o idioma desejado e o diretório de destino onde os dados serão armazenados para serem utilizados nas próximas etapas do processo, como o pré-processamento e a análise de sentimentos.

Para coletar os *tweets* foi utilizada a API gratuita Twitter4j, disponibilizada em [Yamamoto 2008]. Nessa ferramenta é possível capturar o conteúdo dos textos publicados na rede, os usuários que estão publicando, assim como a data e localização de cada publicação. É possível, inclusive, pesquisar por termos específicos, *hashtags* ou usuários, retornando assim *tweets* de acordo com o termo informado na busca. Desta forma, foram coletadas as informações dos *tweets* mais recentes de acordo com o termo de busca informado pelo usuário. Após capturar os *tweets*, foi construído um arquivo para cada *tweet* na API TwitterAPI, contendo o texto da publicação, data e hora, avatar do usuário, quantidade de *retweets* e favoritos. Os arquivos são armazenados no diretório de saída passado como parâmetro para a API. Para fins de identificação, o nome dos arquivos foram compostos pelo nome do usuário juntamente com a data e hora da publicação.

4.3. Pré-processamento

A qualidade da base de dados é um item crucial para o sucesso da Mineração de Textos e Análise de Sentimentos, de forma que é necessário realizar alguns tratamentos para limpar, organizar e padronizar documentos que originalmente são não-estruturados, para que os mesmos possam ser submetidos à análise de sentimentos ou aos algoritmos de mineração de textos. Nesta seção são abordadas as técnicas que foram utilizadas para pré-processar os documentos textuais, a fim de tornar a classificação mais promissora e computacionalmente menos custosa.

Padronização da Coleção Textual. Uma prática comum para padronizar os textos é a remoção de caracteres irrelevantes, como sinais de pontuação, caracteres especiais e alfanuméricos, assim como padronização dos textos em letras minúsculas, evitando que uma mesma palavra seja indexada de formas diferentes, como “Amor” e “amor”, devido aos valores inteiros que representam as palavras serem considerados diferentes por conta do caractere inicial. No dicionário léxico utilizado nesta aplicação as palavras estão padronizadas em letras minúsculas e sem sinais de pontuação, de forma que a etapa de padronização é essencial para o bom funcionamento da análise léxica.

No entanto, considerando que os textos foram extraídos de redes sociais e podem conter *emoticons*, realizar a remoção de alfanuméricos e sinais de pontuação pode acabar removendo também os *emoticons* presentes no texto, os quais são importantes para a análise de sentimentos. Sendo assim, foi construída uma expressão regular para remover alfanuméricos e sinais de pontuação isolados, como os caracteres “!” e “9”, mas manter cadeias de caracteres, pois estas possivelmente podem representar *emoticons*, como “:)” e “:9”.

Além disso, para diminuir a variação ortográfica dos textos, também foram removidas as cadeias de caracteres que possuem repetição, de forma que tais cadeias repetidas são substituídas por apenas dois caracteres representativos, de forma que o termo “kkkkkk” seja substituído pelo termo “kk”.

Na Tabela 1 é possível visualizar um exemplo prático da padronização dos textos.

Tabela 1. Exemplo de Padronização de Textos.

Documento Original	Documento Padronizado
Nossa, queria MUITO 1 açaí com leite condensado!!!!!! kkkkkkk :P	nossa queria muito açaí com leite condensado k :P

Remoção de Stopwords. Uma vez que a base de dados está padronizada em um formato pertinente, outra técnica que pode ser empregada é a remoção de *stopwords*, isto é, a remoção de palavras pouco discriminativas, como preposições e artigos, visto que estas palavras não expressam nenhum sentimento e ocorrem muitas vezes na maioria dos documentos. Como a proposta do artigo é realizar a análise de sentimentos para documentos em inglês e português, a remoção de *stopwords* também precisa levar em consideração a linguagem dos textos.

Neste contexto, foram utilizadas duas listas de *stopwords* (uma em inglês e outra em português), que contém as palavras consideradas irrelevantes, fornecidas na ferramenta PreText [Soares et al. 2008]. Tendo em vista que as listas de *stopwords* contém palavras em uma linguagem formal, as listas foram personalizadas para considerar abreviações e gírias, além de variações ortográficas comuns em redes sociais, como "é" e "eh", uma vez esses termos são frequentes em bases extraídas de redes sociais.

Vale ressaltar que a remoção de *stopwords* é um processo muito importante para a aplicação, visto que em uma eventual exibição de termos mais frequentes os termos conhecidos como *stopwords* seriam eleitos como mais frequentes, embora não agreguem muito valor para a análise, enquanto os termos realmente importantes tornariam-se pouco frequentes em relação aos demais.

Na Tabela 2 é possível visualizar o resultado da remoção de *stopwords*.

Tabela 2. Exemplo de padronização e remoção de stopwords.

Documento Original	Documento Pré-processado
@leandraleal Vendo a segunda temporada de StrangerThings a conta-gotas pra não terminar logo. Amando!!!	leandraleal vendo segunda temporada strangerthings conta-gotas terminar amando

4.4. Análise de Sentimentos

A partir dos documentos pré-processados e padronizados, a base de dados foi submetida ao processo de detecção de sentimentos, a fim rotular os *tweets* entre as classes positiva, negativa e neutra. Para realizar esta etapa foram utilizadas duas abordagens: (i) aprendizado não-supervisionado, fazendo uso de um dicionário léxico, onde os documentos foram classificados de acordo com a polaridade majoritária das palavras contidas no texto; e (ii) aprendizado semissupervisionado, onde os documentos foram submetidos ao aprendizado transdutivo a partir de um conjunto de treino previamente rotulado. Nesta seção são apresentadas as duas abordagens utilizadas neste processo com maior detalhamento.

4.4.1. Abordagem Não Supervisionada Baseada em Dicionário Léxico

Uma alternativa para analisar sentimentos a partir de textos é fazendo uso de um dicionário léxico, onde é construída uma lista de palavras do idioma desejado que expressam algum tipo de emoção, como as palavras "amor", "feliz" e "adorei" expressam sentimentos positivos, enquanto as palavras "odiei", "nojo" e "péssimo" costumam ocorrer em comentários negativos. Abordagens baseadas em léxico assumem que as palavras individuais possuem uma polaridade prévia, que é uma orientação semântica independente de contexto e que pode ser expressada com um valor numérico ou uma classe [Taboada et al. 2011], conforme o exemplo apresentado na Tabela 3.

Tabela 3. Exemplo de Dicionário Léxico.

Palavras	Classe
gostei, legal, bom	Positivo
chato, ruim, cansado	Negativo
dormi, olhei, pensando	Neutro

Desta forma, foi carregado o dicionário léxico OpenWordNet-PT, disponibilizado em [Rademaker et al. 2010], contendo as palavras do idioma desejado e suas respectivas polaridades, que indicam a propensão de determinada palavra pertencer às classes positiva, negativa ou neutra. Em seguida, foram percorridas todas as palavras do texto e verificado se o dicionário construído contém a palavra e qual é o sentimento dela.

Além de analisar todas as palavras do texto, é necessário analisar também os *emoticons*, pois eles possuem uma alta capacidade de expressar os sentimentos dos usuários, especialmente em redes sociais como o *Twitter*, onde há uma limitação de caracteres e os usuários acabam optando por utilizar tais *emoticons* para expressar sentimentos de forma mais resumida. A maioria dos *emoticons* são baseados em faces humanas e expressam emoções de felicidade ou tristeza, no entanto, nos últimos anos foram surgindo *emoticons* para diversas situações. Desta forma, considerar também os *emoticons* presentes no texto pode ser um recurso útil para a análise de sentimentos, uma vez que *tweets* que contém muitos *emoticons* costumam conter mais elementos significativos do que um texto sem nenhum *emoticon* presente.

Assim como o dicionário léxico, foi construído neste projeto um dicionário de *emoticons*, contendo cada *emoticon* e sua respectiva classe. Na Tabela 4, são apresentados alguns exemplos de *emoticons* e suas polaridades.

Tabela 4. Lista de *emoticons*.

Emoticons	Classe
😊 ❤️ 🎉 😍 😊 😊	Positivo
😢 😞 😞 😞 😢 😢	Negativo
😴 😴 😴 😴 😴 😴	Neutro

Para a construção da lista de polaridade de *emoticons*, foram selecionados alguns

emoticons mais representativos para cada classe. As polaridades foram definidas manualmente, devido ao fato de não ter sido encontrado nenhum dicionário muito abrangente, visto que a maioria dos dicionários encontrados consideram apenas os *emoticons* mais simples. Ao todo foram selecionados 144 *emoticons* para o dicionário léxico proposto, sendo este válido para ambos os idiomas.

4.4.2. Abordagem Semissupervisionada Baseada em Aprendizado de Máquina

Dependendo do texto analisado, a abordagem baseada em dicionário léxico pode acabar não apresentando um resultado muito satisfatório, devido ao fato de considerar apenas as palavras que ocorrem em tal dicionário. Neste contexto, foi implementada a técnica de aprendizado transdutivo semissupervisionado, a qual realiza a rotulação automática de textos não rotulados dado um conjunto de textos já rotulados previamente e um conjunto de textos não rotulados [Rossi 2016].

Devido a necessidade de um conjunto de treino com exemplos já rotulados, essa abordagem só pode ser utilizada após o usuário refinar a classificação baseada em dicionário léxico, sendo esta a primeira abordagem utilizada quando não há nenhum representante para as classes. Desta forma, o usuário deve reclassificar os comentários que foram classificados a partir da análise baseada em dicionário léxico, rotulando manualmente os *tweets* dentro das classes positivo, negativo ou neutro. Feito isso, é executada a API TextCategorizationTools, desenvolvida em [Rossi 2016], que irá receber como parâmetro de entrada um diretório de treino, que contém todos os *tweets* que foram rotulados manualmente e o tipo de aprendizado desejado. Além disso, a API recebe como parâmetro a saída desejada, podendo ser um arquivo XML ou apenas apresentando os resultados na tela, informando qual é a confiança de um determinado *tweet* pertencer a uma das classes definidas. A saída utilizada nesta aplicação foi um arquivo no formato XML, para facilitar a apresentação dos resultados posteriormente. Um exemplo do arquivo XML gerado pela API pode ser apresentado na Figura 9.

```
<?xml version="1.0" encoding="utf-8"?>
<documents>
    <configuration learning="semi-supervised" type="classification" language="portuguese"/>
    <document path="emotionmonitor\documento1.txt">
        <class name="positivo" confidence="5.782074508315622"/>
        <class name="negativo" confidence="0.0677"/>
        <class name="neutro" confidence="0.02758"/>
    </document>
    <document path="emotionmonitor\documento2.txt">
        <class name="positivo" confidence="0.0109"/>
        <class name="negativo" confidence="3.65236972735115"/>
        <class name="neutro" confidence="0.01583"/>
    </document>
    ...
</documents>
```

Figura 9. Arquivo XML gerado pela API.

Na Figura 10, é possível visualizar a *interface* desenvolvida para permitir o refinamento dos resultados.

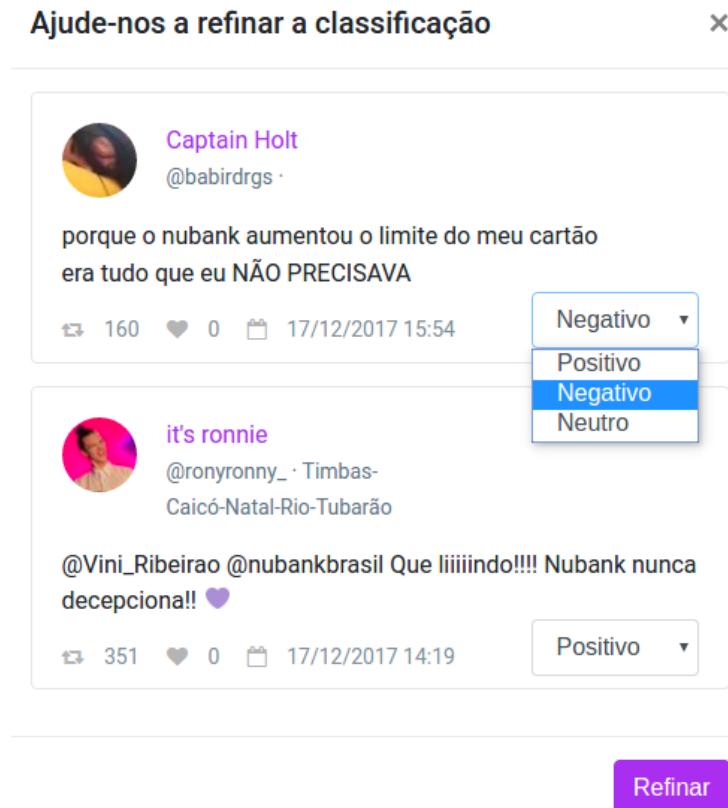


Figura 10. Interface para refinamento dos tweets coletados.

Após o usuário reclassificar os tweets apresentados nas classes corretas, estes tweets são armazenados em pastas para serem utilizados como entrada para o aprendizado transdutivo semissupervisionado, conforme apresentado na Figura 11.



Figura 11. Estrutura de pastas para armazenar o conjunto de treino.

Desta forma, se algum dos *tweets* apresentados não forem rotulados manualmente pelo usuário, eles são inseridos na pasta *unlabeled* para serem reclassificados a partir do conjunto de treino construído. Caso hajam novas consultas e as respectivas pastas já contenham *tweets* representativos, a classificação é realizada exclusivamente pelo aprendizado semissupervisionado.

4.5. Apresentação do Conhecimento

Após realizar as etapas de coleta, onde foram selecionados os dados de interesse, pré-processamento e análise de sentimentos, onde estes dados foram padronizados e os sentimentos expressos no texto foram identificados, o conhecimento extraído da análise pode ser apresentado ao usuário.

Nesta seção são apresentadas algumas funcionalidades desenvolvidas na ferramenta proposta para apresentar o conhecimento obtido.

Overview. Para facilitar o entendimento do usuário, foi desenvolvida uma seção de *overview*, a fim de apresentar de forma mais resumida a polaridade majoritária do comentários analisados, bem como as legendas necessárias para a interpretação dos demais resultados. Dado os trabalhos relacionados apresentados na Seção 3, pode-se observar que a maioria das ferramentas apresentadas disponibilizam uma infinidade de estatísticas, porém não apresentam em nenhum momento um resumo do resultado obtido, de forma que é necessário analisar todas as estatísticas para compreender a polaridade das opiniões expressas pelos usuários. Na Figura 12 pode-se visualizar um exemplo de *overview* gerado pela ferramenta proposta.

Overview

O que estão dizendo sobre o termo nubank?

Pelo que conseguimos analisar, as pessoas estão comentando coisas positivas as respeito disso!

[Entenda nossa análise](#)

Legendas: ● Comentários positivos ● Comentários negativos ● Comentários neutros

Figura 12. *Overview* dos resultados.

Gráfico de Polaridades. Para apresentar um comparativo das polaridades encontradas no texto, foi construído um gráfico de polaridades para apresentar as classes e a quantidade de documentos que foram rotulados em cada uma destas classes, conforme demonstrado na Figura 13.

Gráfico de Polaridades

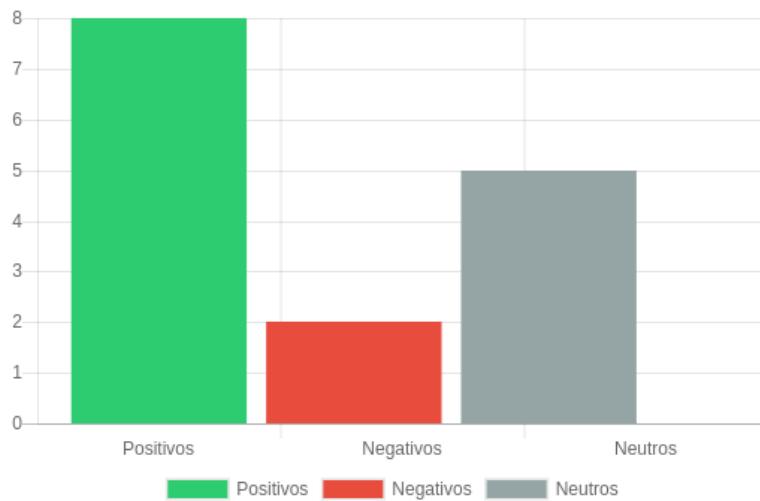


Figura 13. Exemplo de gráfico de polaridades.

Evolução ao Longo do Tempo. Embora o gráfico de polaridades permita identificar a quantidade de documentos classificados em cada classe, essa visualização não permite identificar o comportamento do termo ao longo do tempo. Desta forma, foi construído uma série temporal para apresentar a evolução das classes ao longo do tempo, utilizando o histórico dos exemplos já rotulados anteriormente, conforme a Figura 14.

Evolução nos últimos dias



Figura 14. Exemplo de gráfico de evolução por tempo.

Lista de Tweets Coletados. Também é apresentada a lista de *tweets* que foram coletados e analisados durante o processo, bem como a polaridade expressa em cada

um, conforme ilustrado na Figura 15. Para cada *tweet* são apresentadas algumas informações úteis, como o nome, avatar e *login* usuário que realizou a publicação, data e horário, a quantidade de *retweets* e favoritos da publicação, bem como a polaridade do comentário.



Figura 15. Lista de tweets coletados e analisados.

Nuvem de Termos Frequentes. Também foram coletadas as palavras mais frequentes de todos os documentos para construir uma nuvem de palavras com os termos que estão relacionadas com o termo de interesse, a fim de apresentar ao usuário o que as pessoas estão dizendo sobre o tema, conforme a Figura 16.



Figura 16. Nuvem de termos mais frequentes.

Nuvem de hashtags frequentes. Além dos termos frequentes, também foram coletadas as hashtags que mais ocorrem nos documentos, apresentando ao usuário quais campanhas estão relacionadas ao assunto, conforme apresentado na Figura 17.

Hashtags Frequentes

#virourotina
#imagemdoano
#falaramdemais
#aprocuradeum
#gremio
#mundialdeclubes

Figura 17. Nuvem de hashtags mais frequentes.

5. Estudo de Caso

Dada as funcionalidades desenvolvidas na aplicação, nesta seção é apresentado um estudo de caso a partir do termo de busca "Grêmio" no idioma português, a fim de demonstrar o funcionamento da aplicação.

Após o usuário informar o termo de busca desejado na tela inicial da aplicação, conforme apresentado na Figura 8, a API TwitterAPI desenvolvida neste projeto é executada, recebendo como parâmetro o termo desejado e o diretório onde são armazenados os tweets coletados.

Após a coleta e pré-processamento dos textos, os resultados são transferidos para a aplicação web no formato JSON para serem apresentados ao usuário.

Na Figura 18 pode-se encontrar o resultado obtido na tela de *Overview*.

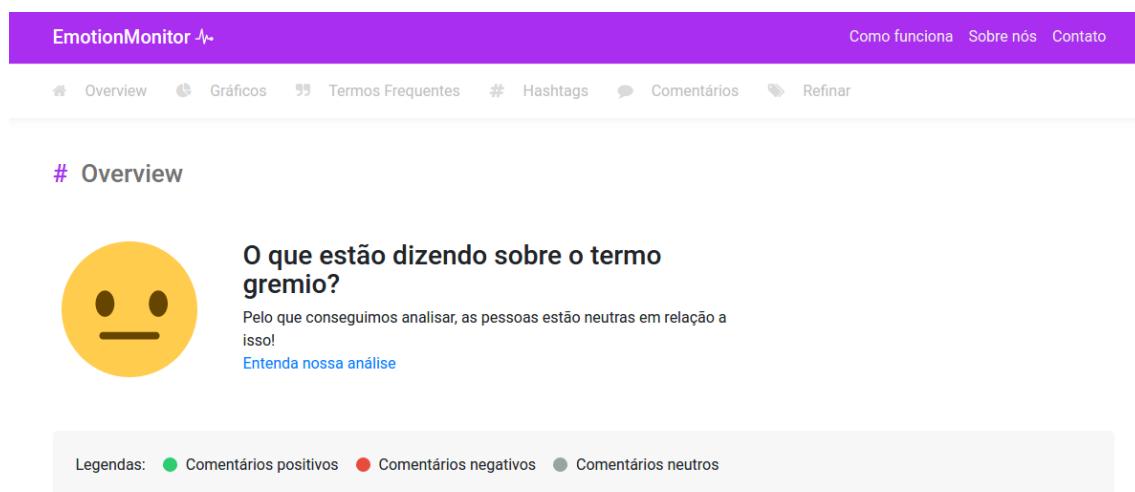


Figura 18. Overview para o termo de busca "Grêmio".

Assim como o *Overview*, também pode-se consultar outros resultados, como os gráficos de polaridade e evolução do termo ao longo do tempo. Na Figura 19, é possível encontrar os gráficos apresentados.



Figura 19. Gráficos para termo de busca "Grêmio".

Além dos gráficos, também é possível visualizar os termos e *hashtags* frequentes, juntamente com a lista de *tweets* coletados, conforme apresentado na Figura 20.

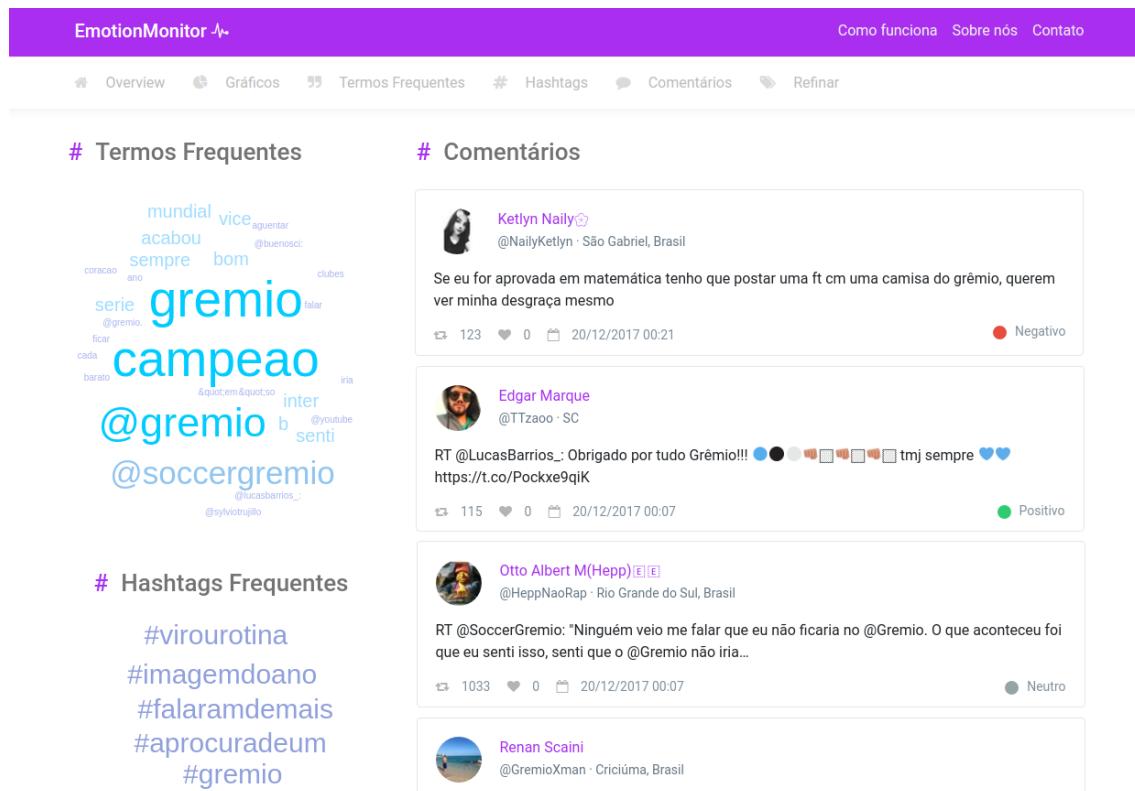


Figura 20. Resultados para o termo de busca "Grêmio".

6. Considerações Finais

O presente trabalho apresenta como contribuição a proposta de uma ferramenta *online*, gratuita e independente de domínio para coletar e analisar textos publicados no *Twitter*, apresentando informações úteis extraídas da base de dados analisada, como a polaridade dos documentos, termos e *hashtags* frequentes no texto, além de oferecer um ambiente de interação com o usuário, de forma que o mesmo seja capaz de refinar os resultados, caso julgue necessário.

Dado o escopo da pesquisa, foram utilizadas algumas técnicas de Mineração de Textos, Análise de Sentimentos e Aprendizado de Máquina para viabilizar o desenvolvimento do projeto, como o pré-processamento dos textos e análise dos sentimentos expressos nos documentos.

Considerando que os dados extraídos do *Twitter* são não-estruturados, foram encontrados alguns desafios durante as etapas do processo de Análise de Sentimentos desenvolvido, destacando-se:

- **Textos informais:** devido ao aspecto espontâneo das publicações e ao limite de caracteres estabelecido, a ortografia dos *tweets* costuma ser mais informal, podendo conter gírias, abreviações e variações ortográficas. Ao construir um dicionário léxico, é necessário considerar todas essas palavras e suas possíveis variações, tornando essa tarefa bastante desafiadora e podendo comprometer diretamente o resultado da detecção de sentimentos, uma vez que tenham sido escondidas palavras pouco abrangentes para compor o dicionário. Na classificação por Aprendizado de Máquina este problema também pode ser observado, pois visto que devido a grande quantidade de variações ortográficas, pode ocorrer de um texto não apresentar nenhuma similaridade com o conjunto de treino, mesmo após a aplicação de várias técnicas de pré-processamento.
- **Tamanho dos textos:** embora os textos publicados no *Twitter* tenham um limite de 140 caracteres, alguns usuários acabam sendo ainda mais sucintos, publicando textos com pouquíssimas palavras. Devido ao fato de a parcela de termos que pode ser consultada no dicionário léxico ou presente no conjunto de treino ser baixa, a classificação pode acabar não encontrando nenhum resultado satisfatório.

Ao longo do desenvolvimento deste projeto de conclusão de curso foram identificados alguns pontos de melhoria e sugestões de trabalhos futuros para aperfeiçoar a ferramenta desenvolvida, destacando-se:

- **Integração com outras redes sociais:** atualmente a ferramenta utiliza apenas documentos extraídos do *Twitter*. No entanto, foi observado que outras redes sociais também poderiam compor a base de dados, como o *Facebook*³, *Instagram*⁴ e *Youtube*⁵, devido a grande quantidade de usuários que utilizam ambas as redes sociais, aumentando assim a abrangência da seleção de dados.
- **Autenticação:** a ferramenta desenvolvida é de consulta pública, não sendo necessário realizar nenhum tipo de autenticação para utilizar os recursos disponíveis. No entanto, foi observado que se houvesse a opção de realizar autenticação, as

³Facebook: www.facebook.com

⁴Instagram: www.instagram.com

⁵Youtube: www.youtube.com

buscas poderiam ser personalizadas de acordo com o perfil do usuário, além de manter um histórico das pesquisas realizadas pelo mesmo. Desta forma, como um trabalho futuro pode-se destacar a autenticação dos usuários, de preferência utilizando as APIs de autenticação das redes sociais utilizadas na coleta dos dados.

- **Revisão e Disponibilização do Dicionário de Emoticons:** o dicionário de *emoticons* construído neste projeto contém 144 *emoticons*, no entanto, a maioria dos exemplos são representativos da classe positiva. Desta forma, como um trabalho futuro, pretende-se coletar mais exemplos para as demais classes, construindo assim um dicionário mais abrangente para ser disponibilizado na ferramenta.
- **Classificação em Tweets Opinativos e não Opinativos:** a classificação de textos realizada nesta ferramenta considera todos os *tweets* como opinativos, isto é, que expressam alguma opinião ou sentimento. No entanto, percebeu-se que alguns dos textos publicados no *Twitter* são de caráter não opinativo, isto é, que carregam apenas fatos e nenhuma opinião, como *retweets* de publicações de jornais e revistas. Como um trabalho futuro, pretende-se fazer a discriminação em *tweets* opinativos e não opinativos para refinar ainda mais os resultados.
- **Representação de Textos:** embora o método baseado em Aprendizado de Máquina tenha apresentado um desempenho satisfatório, como um trabalho futuro pretende-se testar outras técnicas de representação de textos para melhorar a qualidade dos resultados extraídos por Aprendizado de Máquina.
- **Combinação de Léxico e Aprendizado de Máquina:** ao longo do desenvolvimento da pesquisa, foram testados os métodos baseado em dicionário léxico e Aprendizado de Máquina com diversos termos de busca. Na maioria dos testes, o dicionário léxico teve um desempenho inferior ao Aprendizado de Máquina, dado a dificuldade para construir um dicionário abrangente. No entanto, em alguns casos o Aprendizado de Máquina também apresentou resultados inconsistentes, devido a pouca quantidade de *tweets* representativos que foram rotulados manualmente para cada classe. Desta forma, como um trabalho futuro, pretende-se realizar a combinação dos dois métodos para obter um melhor desempenho.

Referências

- Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.
- Aranha, C. and Passos, E. (2006). A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.21529/RESI*, 5(2).
- Araújo, M., Gonçalves, P., Cha, M., and Benevenuto, F. (2014). ifeel: a system that compares and combines sentiment analysis methods. <http://blackbird.dcc.ufmg.br:1210/>. Acessado em 19 de Dezembro de 2017.
- Chen, H. (2001). *Knowledge management systems: a text mining perspective*. Knowledge Computing Corporation.
- Jonn, J. (2008). Socialmention. <http://socialmention.com/>. Acessado em 19 de Dezembro de 2017.
- Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

- MacLennan, Ferranty, M. L., Fabris Lugoboni, L., Moreira Zittei, M. V., Yassuo Tabata, R., and Luiz Correa, H. (2014). Associação entre intensidade de uso de mídias sociais, credibilidade e decisão de compra. *NAVUS-Revista de Gestão e Tecnologia*, 4(2).
- Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.
- Ortony, A., Clore, G. L., and Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.
- Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.
- Rademaker, A., Paiva, V., Freitas, C., Real, L., and Chalub, F. (2010). Openwordnet-pt. <http://wnpt.brlcloud.com/wn/>. Acessado em 19 de Dezembro de 2017.
- Rios, R. A., Pagliosa, P. A., Ishii, R. P., and de Mello, R. F. (2017). Tsviz: a data stream architecture to online collect, analyze, and visualize tweets. www.tsviz.com.br. Acessado em 19 de Dezembro de 2017.
- Rodrigues, C. A. S., Vieira, L. L., Malagoli, L., and Timmermann, N. (2010). Mineração de opinião/análise de sentimentos. *Trabalho acadêmico, Universidade Federal de Santa Catarina, Florianópolis*. www.inf.ufsc.br/~alvares/INE5644/MineracaoOpiniao.pdf.
- Rossi, R. G. (2016). *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Universidade de São Paulo.
- Silva, R. S. (2016). Sistema para identificação de sentimentos em textos da web. www.analisedesentimento.com.br. Acessado em 19 de Dezembro de 2017.
- Soares, M. V. B., Prati, R. C., and Monard, M. C. (2008). *Pretext II: Descrição da reestruturação da ferramenta de pré-processamento de textos*. ICMC-USP.
- Sprinklr (2009). Scup.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.
- Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, volume 8, pages 65–70. sn.
- Ur-Rahman, N. and Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, 39(5):4729–4739.
- Viragine, L., Viragine, I., and Viragine, G. (2010). v-tracker - monitoramento de redes sociais.
- Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.
- Yamamoto, Y. (2008). Twitter4j. www.twitter4j.org/en/index.html. Acessado em 19 de Dezembro de 2017.