



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ESCOLA DE INFORMÁTICA APLICADA

Análise dos Tweets sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos

Wilian Pereira da Silva Santos

Orientador

Flávia Maria Santoro

João Carlos de Almeida Rodrigues Gonçalves

RIO DE JANEIRO, RJ – BRASIL

JANEIRO DE 2016

Análise dos Tweets sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos

Wilian Pereira da Silva Santos

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada por:

Flávia Maria Santoro

Fernanda Araújo Baião

Kate Cerqueira Revoredo

RIO DE JANEIRO, RJ – BRASIL.

JANEIRO DE 2016

Agradecimentos

Agradeço, primeiramente, a Deus por ter me dado saúde e motivação para que eu pudesse chegar até aqui. Agradeço também a toda a minha família, especialmente meus pais, Abel e Marli, por acreditarem em mim, pela dedicação e apoio de sempre. Agradeço a minha avó, Jocelina, e ao meu tio e padrinho, Adilson, por também acreditarem e me incentivarem desde pequeno.

Agradeço aos meus orientadores, Flávia e João, por toda a paciência, dedicação e apoio neste trabalho. Agradeço também a todo o corpo docente da Unirio que me transmitiu todo o conhecimento necessário para que eu conseguisse chegar até aqui.

Por fim, agradeço a todos os meus amigos que estão comigo desde o início da faculdade por todos os momentos que passamos e por toda a ajuda e conhecimento compartilhado durante a graduação.

RESUMO

O crescimento das redes sociais nos últimos anos, bem como a necessidade de seus usuários de expressarem cada vez mais suas opiniões, fez com que muitas organizações prestassem mais atenção no conteúdo gerado pelas mesmas, a fim de utilizarem essas opiniões para adquirirem vantagem competitiva em relação a seus concorrentes. O Twitter é uma dessas redes sociais, principalmente pela facilidade com que as pessoas têm de demonstrar suas opiniões no limite de 140 caracteres, permitindo que sejam mais objetivas. Porém, para que essas opiniões sejam transformadas em informações relevantes que ajudem as empresas a atingirem seus objetivos, é necessária uma solução que realize essas tarefas de maneira automatizada. A Mineração de Textos é um dos processos mais utilizados para a descoberta de novas informações, previamente desconhecidas, através da extração automática dessas informações em diferentes tipos de textos. Juntamente com a Mineração de Textos, a Análise de Sentimentos permite analisar opiniões e seus sentimentos, principalmente através da polaridade, analisando, por exemplo, se o que os clientes de uma marca estão comentando sobre ela nas redes sociais é positivo ou negativo. O presente trabalho analisou o sentimento das opiniões sobre a Black Friday quanto a suas polaridades. O objetivo foi verificar se os comentários das pessoas sobre o evento eram positivos ou negativos. Para isso, foram coletadas do Twitter as opiniões dos usuários e utilizadas técnicas de Mineração de Texto, Análise de Sentimentos e Processamento de Linguagem Natural para a extração de informações relevantes sobre a Black Friday. Ao final, os resultados foram analisados e, em alguns casos, comparados com outras pesquisas. Foi concluído que, apesar dos resultados não poderem ser encarados como verdade absoluta, a solução foi capaz de identificar opiniões positivas e negativas acerca da Black Friday, atingindo, assim, o objetivo inicial.

Palavras-chave: Mineração de Texto, Análise de Sentimentos, Processamento de Linguagem Natural, Twitter, BlackFriday.

ABSTRACT

The growth of social networks in recent years, and the need of its users to express more and more their opinions, has caused many organizations to pay more attention to the content generated by it, in order to use these opinions to gain competitive advantage over their competitors. Twitter is one of those social networks, especially due to the ease with which people have to show their opinions in the 140-character limit, allowing them to be more objective. However, to transform these opinions into relevant information that help companies achieve their goals, a solution that performs these tasks in an automated manner is required. Text Mining is one of the most commonly used processes for the discovery of new information, previously unknown, by automatically extracting this information in different types of texts. Along with Text Mining, Sentiment Analysis allows organizations to analyze opinions and its sentiments, mainly through polarity, considering, for example, if what the customers of a brand say about it on social networks is positive or negative. This research analyzed the sentiment of the Black Friday opinions by their polarity. The goal was to determine whether people's comments on the event were positive or negative. For this reason, the opinions of Twitter users were collected and applied Text Mining, Sentiment Analysis and Natural Language Processing techniques to extract relevant information on Black Friday. At the end, the results were analyzed and, in some cases, compared with other researches. It was concluded that, although the results cannot be regarded as absolute, the solution was able to identify positive and negative opinions on Black Friday, reaching thus the initial goal.

Keywords: Text Mining, Sentiment Analysis, Natural Language Processing, Twitter, Black Friday.

Índice

1	Introdução	9
1.1	Motivação	9
1.2	Objetivos	10
1.3	Organização do texto	11
2	Revisão Bibliográfica	12
2.1	Mineração de texto.....	12
2.1.1	Processo de Descoberta de Conhecimento em Textos	13
2.1.1.1	Coleta.....	14
2.1.1.2	Pré-Processamento	14
2.1.1.3	Mineração	17
2.1.1.4	Análise	18
2.2	Análise de sentimentos	18
2.2.1	Abordagem com Aprendizado de Máquina	20
2.2.1.1	Aprendizado supervisionado	20
2.2.2	Naïve Bayes.....	21
2.3	Twitter.....	22
2.4	Black Friday.....	24
3	Aplicação no domínio.....	25
3.1	Coleta	25
3.2	Pré-processamento	26
3.2.1	Remoção de tweets repetidos	26
3.2.2	Transformação de tweets em formato Unicode.....	27
3.2.3	Transformação de tweets para letras minúsculas	27
3.2.4	Transformação de acrônimos e abreviações de internet.....	27
3.2.5	Tokenização	29
3.2.5.1	Geração simples de tokens	29
3.2.5.2	Identificação de Abreviações.....	30
3.2.5.3	Identificação de Palavras Combinadas	31
3.2.5.4	Identificação de Símbolos de Internet e Números.....	31

3.2.6 Remoção de stopwords, pontuações e caracteres especiais	32
3.3 Mineração de texto e análise de sentimentos	33
3.3.1 Treinamento do corpus.....	34
3.3.2 Análise de sentimento	36
4 Análise dos resultados	40
5 Conclusão	46
5.1 Trabalhos futuros	47

Lista de Figuras

Figura 1 – Processo de Descoberta de Conhecimento em Textos.....	13
Figura 2 – “Linha de Montagem” de um processo de tokenização.....	16
Figura 3 – Texto após passar pelo processo de remoção de stop words	17
Figura 4 – Técnicas de classificação de sentimento	19
Figura 5 – Teorema de Bayes adaptado	21
Figura 6 – Trending Topics no Twitter	23
Figura 7 – Parte de um arquivo json com alguns campos de um tweet.....	26
Figura 8 – Tweets com caracteres Unicode.....	27
Figura 9 – Exemplo de tweet após passar pelo processo de tokenização.....	30
Figura 10 – Gráfico com o sentimento dos tweets da Black Friday.....	41
Figura 11 – Nuvem de palavras com as palavras mais frequentes dos tweets	42

Lista de Tabelas

Tabela 1 – Principais acrônimos e abreviações de internet encontrados no tweets	28
Tabela 2 – Termos em inglês na forma contraída	30
Tabela 3 – Lista com as principais abreviações dos tweets.....	31
Tabela 4 – Lista de stop words da API NLTK 3.0.4	32
Tabela 5 – Lista de pontuações e caracteres especiais	33
Tabela 6 – Resumo das medidas utilizadas para validar o classificador	36
Tabela 7 – Comparação entre sentimentos classificados automaticamente e manualmente	36
Tabela 8 – Comportamento dos tweets classificados após novo treinamento	38
Tabela 9 – Marcas mais comentadas nos tweets	43
Tabela 10 – Marcas mais comentadas nos tweets classificados como positivos.....	44
Tabela 11 – Marcas mais comentadas nos tweets classificados como negativos.....	44

1 Introdução

1.1 Motivação

Com o surgimento e a popularização das redes sociais, muitas pessoas começaram a utilizar esses espaços não só para se relacionarem umas com as outras, como também como meio para emitir suas opiniões sobre um assunto. O crescimento do acesso à internet nos smartphones também contribuiu positivamente para isso. No Brasil, a quantidade de pessoas que utilizam a internet no smartphone chegou a 68,4 milhões no primeiro trimestre de 2015 e representa um crescimento de cerca de 10 milhões em relação ao trimestre anterior (Nielsen, 2015).

A facilidade de acesso à internet permitiu que os usuários pudessem acessar suas redes sociais em qualquer lugar e a qualquer hora. Isso contribuiu para que as pessoas pudessem dar suas opiniões sobre um produto ou evento, por exemplo, no contexto em que se encontravam. Dessa forma, as redes sociais ganharam outros objetivos além de manter as pessoas conectadas umas com as outras. Hoje, as opiniões emitidas nesses espaços passaram a ser vistas com outros olhos pelas organizações que perceberam que poderiam ganhar vantagem competitiva com essas informações, utilizando-as como meio para medir a reputação de uma marca ou produto e até mesmo para recomendar certos produtos de acordo com a avaliação feita pelos usuários.

Mas de que modo essas organizações poderiam transformar as opiniões de seus clientes nas redes sociais em informações relevantes para que pudessem ser usadas na tomada de decisão? O primeiro obstáculo está na forma como esses dados estão disponíveis na rede. Por se tratarem de opiniões, essas informações são encontradas na forma de texto livre, ou seja, são dados que não possuem uma estrutura definida, conhecidos como dados não estruturados.

Outra dificuldade é o volume e a velocidade cada vez maior que esses dados são disponibilizados. A cada segundo uma nova opinião é postada nas redes sociais. Como administrar essa enorme quantidade de dados e, ao mesmo tempo, extrair informações relevantes? Para isso, são utilizadas técnicas de mineração de texto para a extração do conhecimento necessário dessas fontes.

A mineração de textos é definida por Hearst (2003) como "o processo de descobrir computacionalmente novas informações, previamente desconhecidas, pela extração automática de informação de diferentes recursos de texto". Porém, o conhecimento propriamente dito só será alcançado mediante uma análise e contextualização dos dados. Diversas técnicas de mineração de texto são aplicadas nos conteúdos extraídos das redes sociais para facilitar essa análise, uma vez que os dados podem vir incompletos, redundantes e com informação irrelevante.

Em alguns casos, também pode ser necessário determinar o sentimento desse conteúdo coletado ao classificar automaticamente a opinião e o sentimento desses textos. Isso é possível através da análise de sentimentos, que é "o campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, estimativas, atitudes e emoções sobre entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos" (LIU, 2012).

No passado, as pessoas precisavam perguntar a opinião do outro sobre determinado produto ou serviço. Hoje, essas opiniões encontram-se espalhadas por toda a Internet. Antes, as organizações precisavam realizar pesquisas de opinião e grupos de discussão para capturar as opiniões de seus consumidores. Atualmente, tornou-se mais fácil para as empresas conhecerem mais as vontades e necessidades dos seus clientes, principalmente, através das redes sociais. Com isso, a quantidade de pesquisas tanto na área de mineração de textos como na área de análise de sentimento cresceu muito desde os anos 2000 (LIU, 2012).

1.2 Objetivos

Este trabalho de conclusão de curso visa compreender o conhecimento do evento conhecido como "Black Friday" no Twitter, ocorrido no dia 27 de novembro de 2015, além de analisar o sentimento das opiniões das pessoas sobre esse evento quanto a polaridade das mesmas, isto é, se as opiniões são positivas ou negativas. Para isso, serão utilizadas técnicas de mineração de texto e análise de sentimentos.

A abordagem consiste em coletar as opiniões dos usuários do Twitter sobre o evento durante a ocorrência do mesmo. Em seguida, serão aplicadas as técnicas de mineração de texto e análise de sentimentos para extrair o conhecimento inerente aos dados analisados. Por último, serão apresentados os resultados da análise.

1.3 Organização do texto

O presente trabalho está estruturado em capítulos e, além desta introdução, está organizado da seguinte forma:

- Capítulo 2: Este capítulo contém a Revisão Bibliográfica do trabalho e apresenta os principais conceitos utilizados relacionados à Mineração de Texto, Análise de Sentimento, Twitter e Black Friday.
- Capítulo 3: Apresenta uma descrição detalhada do processo de Mineração de Texto e Análise de Sentimentos aplicados no domínio da Black Friday.
- Capítulo 4: Este capítulo mostra os resultados obtidos pela Mineração de Texto e Análise de Sentimentos detalhados no capítulo anterior e realiza uma comparação com outras pesquisas relacionadas ao evento.
- Capítulo 5: Finalmente, este capítulo apresenta as considerações finais do trabalho e sugestões para trabalhos futuros.

2 Revisão Bibliográfica

2.1 Mineração de texto

Assim como a Mineração de Dados é uma etapa do processo de Descoberta de Conhecimento em Base de Dados (KDD), a Mineração de Texto é uma etapa do processo de Descoberta de Conhecimento em Textos (KDT) e é definida por Hearst (2003) como "o processo de descobrir computacionalmente novas informações, previamente desconhecidas, pela extração automática de informação de diferentes recursos de texto". Faz uso de técnicas de recuperação de informação, extração de informação e processamento de linguagem natural (PLN) e as conecta com os algoritmos e métodos do KDD (Knowledge Discovery in Databases), mineração de dados, aprendizado de máquina e estatística (HOTHO; NÜRNBERGER; PAAB, 2005).

O grande desafio da mineração de texto é exatamente descobrir informações que ainda não são conhecidas e que serão extraídas de documentos de texto, comentários em redes sociais, etc. Ainda de acordo com Hearst (2003), a mineração de texto é uma variação da mineração de dados que, por sua vez, procura padrões em grandes bancos de dados, onde os dados possuem uma estrutura definida, ao contrário da mineração feita com textos, que utiliza no estudo, geralmente, textos em linguagem natural.

Para Miner et al. (2012, p. 32), a mineração de texto possui sete áreas de atuação, listadas a seguir:

1. Pesquisa e recuperação de informação (IR): armazenamento e recuperação de documentos de texto, incluindo ferramentas de pesquisa e palavras-chave;
2. Clusterização de documentos: agrupamento e categorização de termos, trechos, parágrafos ou documentos, utilizando métodos de clusterização de mineração de dados;
3. Classificação de documentos: agrupamento e categorização de trechos, parágrafos ou documentos usando métodos de classificação de mineração de dados, baseados em modelos treinados em exemplos pré-classificados;
4. Web mining: Mineração de dados e textos na Internet, com foco específico na escala e interconectividade da Web;

5. Extração de informação (IE): Identificação e extração de fatos e relacionamentos relevantes de textos não estruturados. Processo de tornar textos não estruturados e semiestruturados em dados estruturados.
6. Processamento de linguagem natural (NLP): Processamento de linguagem de baixo nível e compreensão de tarefas (por exemplo, tagging part-of-speech). Muitas vezes, usado como sinônimo para linguística computacional;
7. Extração de conceito: Agrupamento de palavras e frases em grupos semanticamente similares.

Classificação de documentos e NLP são as áreas que mais se identificam com o propósito deste trabalho, uma vez que os textos serão extraídos do Twitter e classificados de acordo com a sua polaridade (positivo ou negativo) a partir de exemplos pré-classificados, utilizando também técnicas de processamento de linguagem natural. Além dessas duas áreas, a seleção de características (feature selection), que tem por objetivo reduzir a dimensionalidade do conjunto de dados procurando manter suas características relevantes (CONTRERAS, 2002), também se insere no contexto deste trabalho.

2.1.1 Processo de Descoberta de Conhecimento em Textos

Existem diferentes abordagens no que diz respeito ao processo de Descoberta de Conhecimento em Textos. No entanto, todas estão interessadas em atingir o mesmo objetivo, ou seja, descobrir novas informações que ajudem na tomada de decisões. Aranha (2007) observou os modelos adotados em outros trabalhos da literatura e propôs um modelo de processo com as seguintes etapas: coleta, pré-processamento, indexação, mineração e análise. O modelo proposto por Aranha será adotado neste trabalho. Porém, apenas a etapa de indexação, que permite uma busca com maior rapidez em grandes volumes de textos, não será utilizada, uma vez que não serão realizadas consultas nos textos avaliados. Assim, as etapas para a mineração de texto seguidas neste trabalho estão representadas na Figura 1.



Figura 1 – Processo de Descoberta de Conhecimento em Textos

Fonte: Adaptado pelo autor

2.1.1.1 Coleta

A etapa de coleta é importante, pois é responsável pela extração dos textos que servirão de base para a análise. Essa base pode ser estática, quando os dados não mudam, ou dinâmica, quando os dados podem ser atualizados automaticamente por meio de robôs autônomos. Os textos podem ser extraídos de tabelas de vários bancos de dados, de diretórios de pastas de um HD e da Internet, de acordo com a relevância dos mesmos ao domínio de estudo.

A Internet possibilitou que uma infinidade de textos pudesse ser consultada e servissem de base para diferentes domínios. Artigos, livros, documentos, páginas e comentários são alguns exemplos de conteúdo que estão presentes na web e que podem ser utilizados nessa fase. Existem diversas ferramentas que possibilitam a coleta de textos da web. Uma das formas mais utilizadas para realizar essa coleta é através de web crawlers, que são robôs responsáveis por “varrer” a rede de forma autônoma e são capazes, por exemplo, de identificar em uma página HTML apenas seu conteúdo texto.

Neste trabalho, apesar de utilizar conteúdo presente na web, os textos a serem analisados serão coletados através da API de streaming do Twitter, que retornará os comentários feitos pelos usuários na rede social, bem como outras informações que serão relevantes para a análise, como o idioma.

2.1.1.2 Pré-Processamento

Após a coleta, os textos podem conter caracteres ou termos que não são importantes para a análise. Os textos também precisam ser representados em uma forma mais estruturada, mas sem que a informação presente no texto perca seu sentido. Essa etapa, de modo geral, é responsável pela preparação dos dados para serem processados pela próxima fase.

Uma das técnicas mais utilizadas no pré-processamento é o Processamento de Linguagem Natural (PLN), que, segundo Liddy (2001), é um conjunto de técnicas computacionais que procura analisar e representar dados textuais a fim de alcançar um processamento de uma linguagem similar ao humano para uso em diversas tarefas e aplicações. Entre as principais técnicas de PLN estão a divisão do texto em palavras (tokenização), filtragem, remoção de stop words, stemização, etiquetagem POS (Part of Speech), entre outros. A seguir, serão detalhadas apenas a tokenização e a remoção de stop words, pois serão as únicas técnicas que vão ser utilizadas no processo de Descoberta de Conhecimento em Textos deste trabalho.

a) Tokenização

A tokenização consiste na segmentação de um texto em unidades linguísticas como palavras, pontuações, números, alfanuméricos, etc. (TRIM, 2013) Token é o nome que se dá a essas palavras depois da segmentação. Por exemplo, a frase “O Rio de Janeiro é um estado da região Sudeste do Brasil”, após passar pelo processo de tokenização mais simples, seria segmentada dessa forma:

[O] [Rio] [de] [Janeiro] [é] [um] [estado] [da] [região] [Sudeste] [do] [Brasil.]

Assim como em algumas outras línguas, a língua portuguesa separa suas palavras por meio de espaços em branco. Porém, algumas palavras como “Rio de Janeiro”, por exemplo, somente mantêm seu significado quando são representadas em conjunto. Já na língua inglesa, “I’m”, que é a forma abreviada de “I am”, geralmente, não é separada como sua forma abreviada. Para que o processamento do texto não seja prejudicado nos estágios posteriores, antes de começar o processo de tokenização, muitas vezes, é necessário identificar as unidades que não podem ser decompostas ou como elas deveriam ser decompostas para facilitar o entendimento de acordo com o domínio de estudo.

Muitos autores propuseram algoritmos para tentar solucionar esse problema. Konchady (2006) propôs uma metodologia (Figura 2) para a identificação de tokens e utiliza um dicionário de dados e regras de formação de palavras para preservar a semântica presente nos tokens de um texto antes do processo de tokenização (CARRILHO, 2007).

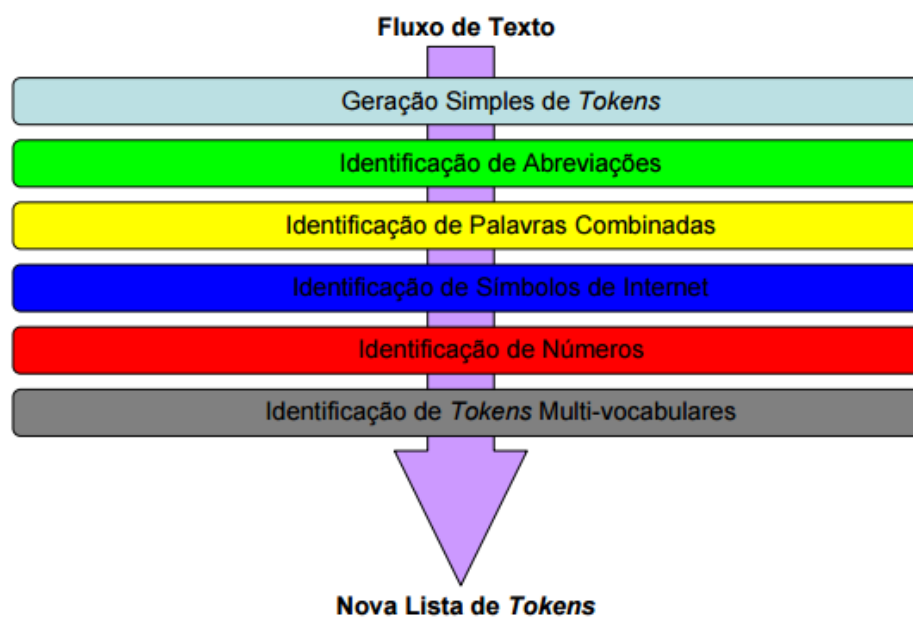


Figura 2 – “Linha de montagem” de um processo de tokenização

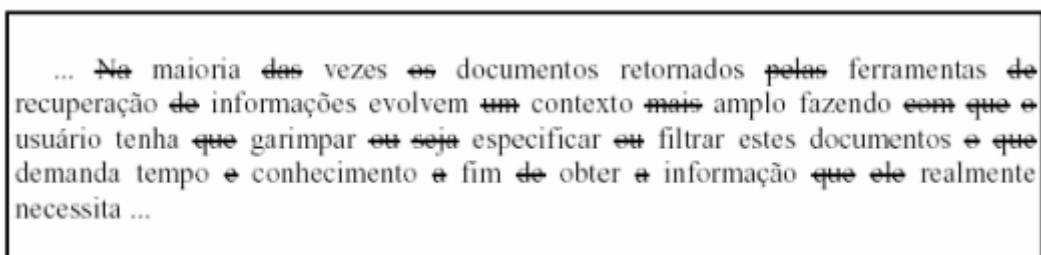
Fonte: Adaptado por Carrilho (2007)

Semelhante a uma “linha de montagem”, o processo começa com o “Fluxo de Texto” que é qualquer sequência de caracteres. Depois, o texto passa pela “Geração Simples de Tokens” onde são identificados, de forma simples, os tokens a partir de uma lista de delimitadores e o espaço em branco. Em “Identificação de Abreviações”, utilizam-se dicionários para identificar as abreviações presentes no texto. Palavras separadas por caracteres como “&” e “-“ são unidas, formando um único token, como “AT & T” na etapa “Identificação de Palavras Combinadas”. Em “Identificação de Símbolos de Internet” são identificados endereços de e-mail, endereços de sites e endereços IP. Em “Identificação de Números” são observadas todas as representações de números, incluindo medidas e valores. Já em “Identificação de Tokens Multi-vocabulares” identificam-se as palavras que precisam estar no mesmo token para que não percam seu sentido original.

b) Remoção de stop words

As stops words são palavras que não são consideradas relevantes em um processo de mineração e que contribuem apenas para a compreensão geral de um texto. O conjunto dessas palavras é conhecido como stoplist. Uma lista de stop words contém as palavras que mais aparecem em um texto como, por exemplo, artigos, preposições, pontuação, conjunções e pronomes de uma língua. (ARANHA, 2007)

Há duas formas de definir uma stoplist: manualmente ou de forma automática. Manualmente, por meio de um especialista no domínio do assunto, ou de forma automática, através da frequência de aparição das palavras. Por exemplo, na frase “O Rio de Janeiro é um estado da região Sudeste do Brasil”, poderiam ser consideradas como stop words as palavras “O”, “de”, “é”, “um”, “da” e “do”. Porém, se a preposição “de” for removida, tiraria o sentido de “Rio de Janeiro”. Casos como esse devem ser analisados com mais cuidado. A eliminação das stop words reduz substancialmente a quantidade de termos e, conseqüentemente, também diminui o custo computacional das próximas etapas. A Figura 3 apresenta um texto após ser validado por uma stop list (Aranha, 2007).



... ~~No~~ maioria ~~das~~ vezes ~~os~~ documentos retornados ~~pelos~~ ferramentas de recuperação ~~de~~ informações envolvem ~~um~~ contexto ~~mais~~ amplo fazendo ~~em~~ que o usuário tenha ~~que~~ garimpar ~~ou~~ seja especificar ~~ou~~ filtrar estes documentos e ~~que~~ demanda tempo e conhecimento a fim de obter a informação ~~que~~ ele realmente necessita ...

Figura 3 – Texto após passar pelo processo de remoção de stop words

Fonte: Aranha (2007)

2.1.1.3 Mineração

A etapa de mineração é a parte central do processo de Descoberta de Conhecimento em Textos e envolve a escolha e aplicação de um ou mais algoritmos adequados para realizar a extração do conhecimento. Esses algoritmos são aplicados no texto pré-processado na fase anterior, onde os dados já foram devidamente tratados e agora apresentam uma estrutura melhor para a realização do processamento principal. O produto final da etapa de pré-processamento será um conjunto de tokens para cada texto que passou pelo processo de tokenização, sem as stop words e pontuações. É a partir desses tokens que, após a aplicação do algoritmo de mineração escolhido, será extraído o conhecimento desejado.

Existem diversos algoritmos que são comumente utilizados na mineração de textos. Porém, a escolha deverá levar em conta o que se quer obter de informação. Se a necessidade de informação do usuário é relacionar documentos, agrupando textos similares, então a técnica a ser escolhida é a Clusterização. Dessa forma, o algoritmo que poderia ser utilizado seria o K-means, por exemplo. Se o objetivo for a associação de textos a classes pré-definidas, a técnica conhecida como Classificação seria a mais adequada e, nesse caso, o algoritmo

Naïve Bayes poderia ser o mais apropriado. A Sumarização, que é a criação automática de resumos, também é outra técnica de mineração (CARRILHO, 2007). Na seção 2.3.2 será detalhado o algoritmo Naïve Bayes, utilizado neste trabalho para a realização da análise de sentimentos nos textos coletados.

2.1.1.4 Análise

A etapa de análise, também chamada por alguns autores de pós-processamento, é o último estágio do processo de mineração de texto e é responsável por avaliar e interpretar os resultados. Para isso, é necessário, primeiramente, que esses resultados sejam refinados, além de deixá-los coerentes, eliminando a redundância e aperfeiçoando o seu grau de clareza. Algumas das técnicas utilizadas nesta etapa incluem a clusterização e a estruturação das informações encontradas em alguma notação específica. (GONÇALVES, 2010).

Um dos objetivos principais desta fase é apresentar ao usuário final os resultados da pesquisa de uma maneira clara e eficiente para que o mesmo possa extrair conhecimento e novas informações do estudo. Assim, podem ser utilizados para a visualização dos resultados gráficos, tabelas e qualquer outro tipo de artifício que ajude o interessado na pesquisa a entender claramente os objetivos da análise. É importante ressaltar que, caso os resultados encontrados nesta fase ainda não sejam suficientes para a realização de uma boa análise, pode ser necessário o retorno às fases anteriores, o que torna o processo de mineração de texto iterativo.

2.2 Análise de sentimentos

Analisar o sentimento de uma série de comentários de uma rede social sobre um determinado assunto, por exemplo, pode ser difícil se depender apenas do esforço humano. Uma organização que quer conhecer se o que os seus clientes estão falando do seu produto é positivo ou negativo pode lidar com milhões de comentários. Para analisar o sentimento presente nesses comentários de uma forma mais eficaz, essas organizações têm utilizado meios automáticos para obterem um conhecimento mais rapidamente.

A análise de sentimentos, também conhecida como mineração de opinião, pode ser definida como “o campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, estimativas, atitudes e emoções sobre entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos” (LIU, 2012). Dessa

forma, o objetivo da análise de sentimentos é encontrar opiniões, identificar os sentimentos nelas expressos e, então, classificar sua polaridade.

Considerada um problema de classificação de textos, a análise de sentimento pode ter diversas abordagens quanto a esse aspecto. A Figura 4 apresenta um esquema com as diferentes técnicas de classificação de sentimento, que pode ser dividida em Abordagem com Aprendizado de Máquina, Abordagem Léxica e Abordagem Híbrida (MEDHAT; HASSAN; KORASHY, 2014).

A Abordagem com Aprendizado de Máquina, como o nome diz, utiliza técnicas de aprendizado de máquina e é dividida em aprendizado supervisionado e aprendizado não supervisionado. A Abordagem Léxica baseia-se numa coleção de termos de sentimentos conhecidos e pré-compilados e é dividida em abordagem baseada no dicionário e abordagem baseada em um corpus, que usa métodos estatísticos ou semânticos para encontrar a polaridade dos sentimentos. Já a Abordagem Híbrida combina os dois tipos de abordagens anteriores. Este trabalho focará apenas na Abordagem com Aprendizado de Máquina, que será utilizada na implementação do domínio estudado.

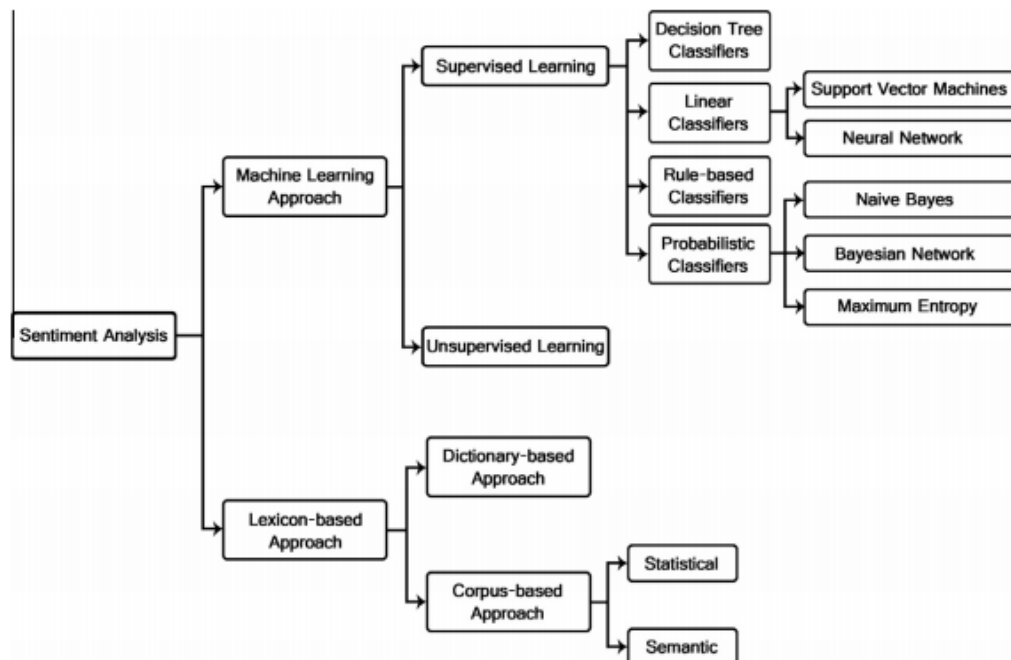


Figura 4 – Técnicas de classificação de sentimento

Fonte: (MEDHAT; HASSAN; KORASHY, 2014)

2.2.1 Abordagem com Aprendizado de Máquina

Aprendizado de máquina pode ser definido como um conjunto de métodos que podem detectar padrões em dados automaticamente e, em seguida, utilizar os padrões descobertos para prever dados futuros ou ainda realizar outros tipos de tomada de decisão sob incerteza (MURPHY, 2012).

A Abordagem com Aprendizado de Máquina envolve os algoritmos de aprendizado de máquina e encara a análise de sentimento como um problema de classificação de texto que faz uso de características linguísticas e/ou sintáticas (MEDHAT; HASSAN; KORASHY, 2014). É dividido em aprendizado supervisionado, que depende de textos rotulados de treinamento, e aprendizado não supervisionado, onde o algoritmo identifica grupos similares sem a necessidade de textos pré-rotulados para treinamento. Este trabalho se aprofundará apenas no aprendizado supervisionado.

2.2.1.1 Aprendizado supervisionado

Como mencionado anteriormente, o aprendizado supervisionado se dá quando os dados de treinamento são rotulados com as respostas corretas aplicadas a um modelo de classificação. Existem vários tipos de classificadores supervisionados:

a) Classificadores Probabilísticos: utilizam inferência estatística para encontrar a melhor classe para um determinado exemplo. Para isso, a probabilidade desse exemplo ser membro de cada uma das possíveis classes é calculada e, então, a classe com maior probabilidade será selecionada como a melhor classe. Os classificadores probabilísticos mais famosos são o Naïve Bayes, Rede Bayesiana e Entropia Máxima (ME).

b) Classificadores Lineares: baseia-se na combinação linear das características dos componentes. Support Vector Machines (SVM) é um exemplo de classificador linear.

c) Classificadores de Árvores de Decisão: organiza uma série de questões e condições na estrutura de árvore. Ao atingir o nó folha, a classe do rótulo associado ao mesmo é atribuída ao registro.

d) Classificadores Baseados em Regras: classifica a partir de um conjunto de regras.

Neste trabalho será utilizado o classificador probabilístico utilizando o Naïve Bayes, abordado no próximo tópico.

2.2.2 Naïve Bayes

O classificador Naïve Bayes é o mais simples. O termo “naive” (ingênuo) é atribuído à independência condicional dos atributos, que quer dizer que a informação de um evento não está relacionada com a informação de outro evento. O Naïve Bayes apresenta ótimos resultados para a categorização de textos (ARANHA, 2007).

Esse classificador utiliza como base o Teorema de Bayes. Para facilitar o entendimento, o teorema foi adaptado para o contexto deste trabalho, utilizando a abordagem de Toit (2015). Dessa forma, para analisar se um texto extraído do Twitter é positivo, por exemplo, a fórmula ficaria conforme a apresentada na Figura 5. A fórmula para verificar se um tweet é negativo seria representada de forma semelhante, apenas trocando os valores do campo positivo pelos valores referentes ao negativo.

$$P(\text{positive}|\text{tweet}) = \frac{P(\text{tweet}|\text{positive}) P(\text{positive})}{P(\text{tweet})}$$

Figura 5 - Teorema de Bayes adaptado

Fonte: Toit, 2015

Nesse caso, a fórmula calcula a probabilidade de um tweet ser classificado como positivo. Para calcular $P(\text{tweet}|\text{positive})$ é necessário um conjunto de tweets treinados que já foram classificados nas duas categorias (positivo e negativo). Assim, pode-se calcular a probabilidade de um tweet ser de uma classe específica.

As chances de encontrar um tweet específico no conjunto de treinamento são relativamente baixas. Dessa forma, o tweet deve ser segmentado em palavras (tokenização) e o cálculo deve ser feito a partir de cada palavra no conjunto de treinamento. A fórmula para calcular se um tweet é positivo, por exemplo, ficaria dessa forma:

$P(\text{tweet}|\text{positive}) = P(T1|\text{positive}) * P(T2|\text{positive}) * \dots * P(Tn|\text{positive}) * P(\text{positive})$,
onde T1, T2... Tn seriam cada palavra do tweet.

Para determinar a probabilidade de uma palavra específica ser de uma categoria precisa-se dividir o número de vezes que T_i ocorre em tweets que estão marcados como positivos num conjunto de treinamento pelo número total de palavras dos tweets que estão

marcados como positivo. Nesse caso, há apenas duas classes possíveis (positivo e negativo), ou seja, $P(\text{positive})$ será 0,5 e indica que há 50% de chances do tweet ser positivo.

Ao invés de calcular a probabilidade de ser um tweet ($P(\text{tweet})$) explicitamente, o algoritmo fará o cálculo do denominador para cada categoria e, então, os normalizará para que a soma dê um:

$$\text{SUM}[\text{positive}](P(\text{positive}) * P(T1|\text{positive}) * ... * P(Tn|\text{positive}))$$

O procedimento acima pode ser usado para calcular a probabilidade das duas classes (positivo e negativo). Após o cálculo da probabilidade de cada classe ser conhecida, ambas são comparadas e a classe que obteve o maior valor de probabilidade é usada como a classe do *tweet*. Por exemplo, após o algoritmo calcular a probabilidade de um determinado *tweet* (conjunto de tokens) ser positivo e a probabilidade do mesmo ser negativo, as probabilidades serão comparadas e a classe que obteve a maior probabilidade será atribuída a esse texto.

Em Aprendizado de Máquina, o classificador Naïve Bayes é comumente utilizado através de APIs que implementam o Teorema de Bayes, como o módulo presente na API NLTK para a linguagem de programação Python (NLTK, 2015).

2.3 Twitter

O Twitter é uma rede social que permite que seus usuários postem suas ideias e compartilhem informações em tempo real. Segundo o Site Oficial da empresa (2015), o Twitter possui cerca de 316 milhões de usuários ativos mensalmente e são postados diariamente 500 milhões de “tweets” (textos enviados pelos usuários), números que comprovam a relevância dessa rede social em relação ao alcance de pessoas e o porquê tem se revelado uma excelente fonte de informações para pesquisas.

Também conhecido como um microblog, o Twitter permite que seus usuários enviem “tweets” de no máximo 140 caracteres, links, fotos, gifs animados e, mais recentemente, vídeos. Os usuários podem seguir outros usuários e, assim, receber em tempo real as atualizações em sua timeline. Uma timeline é exatamente uma linha do tempo em que os “tweets” de um usuário e de seus seguidores são exibidos em ordem cronológica.

Cada usuário pode mandar mensagens ou responder outros usuários publicamente através do “reply”. Mensagens privadas também podem ser enviadas, mas somente para as

pessoas que seguem o usuário, através da funcionalidade “Direct Message” ou simplesmente DM, como é popularmente conhecida.

Um “tweet” também pode ser compartilhado com os seguidores através do “Retweet”. Dessa forma, o post da pessoa aparece na timeline mesmo que os seguidores de determinado usuário não sigam o perfil do “tweet” compartilhado. Semelhante ao “curtir” do Facebook, o Twitter também possui a funcionalidade de marcar como favorito um post.

A fim de categorizar as mensagens postadas no Twitter, os usuários utilizam a hashtag, representada pelo símbolo #. A hashtag geralmente é postada para marcar o(s) assunto(s) relacionado(s) a um “tweet”. Por exemplo, em um “tweet” sobre o lançamento do iPhone 6, um usuário poderia utilizar a hashtag #iPhone6. Ao clicar numa hashtag, uma lista de “tweets” que contêm a mesma marcação aparece em ordem cronológica de postagem. Algumas hashtags que se tornam populares, ou seja, são postadas em vários “tweets”, por diferentes pessoas, em um curto espaço de tempo, aparecem nos “Trending Topics”, que é uma lista ordenada dos assuntos mais comentados do Twitter (Figura 6). Nessa lista, também aparecem termos que não são exatamente hashtags, mas que aparecem com frequência num espaço de tempo na rede social. Embora não seja obrigatória na postagem de um “tweet”, a hashtag é muito utilizada pelos usuários e é um dos meios que facilita a extração de conhecimento no Twitter.

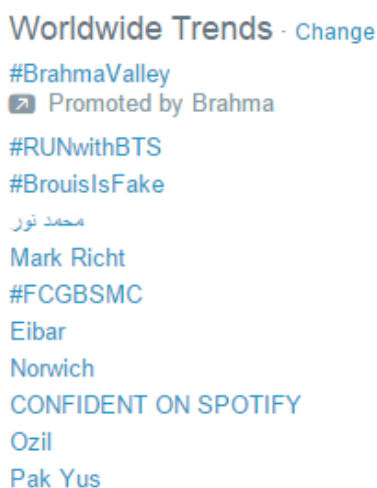


Figura 6 – Trending Topics no Twitter

Fonte: Elaborado pelo autor

Este trabalho utilizará o Twitter para extrair as opiniões dos usuários sobre determinado evento através dos “tweets” postados pelos mesmos e dos termos e hashtags

inseridos no conteúdo de cada “tweet”. Para isso, será utilizado a API de Streaming do Twitter para obter essas informações.

2.4 Black Friday

Nos Estados Unidos, Black Friday refere-se ao dia depois do Dia de Ação de Graças (Thanksgiving), ou seja, na última sexta-feira de novembro (ABOUT, 2015). Originalmente foi chamado de “Black Friday” pelo Departamento de Polícia de Filadélfia, pois a quantidade de pessoas que ia às compras nesse dia era tão grande que causava acidentes de trânsito e até violência. O dia marca oficialmente o início da temporada de compras de Natal nos EUA e chega a oferecer descontos de até 80% em lojas físicas e virtuais do país. A Black Friday também vem acontecendo em outros países como Reino Unido e Brasil. Por aqui, a data acontece desde 2010 (G1, 2010) e tem crescido a cada ano.

No início, os comerciantes não gostavam de associar seus negócios ao termo “Black Friday”, uma vez que “black” foi associado várias vezes na História a fracasso, como em “Black Monday”, que se refere ao dia em que a Dow Jones Average caiu cerca de 22% no mercado de ações, ou em “Black Thursday”, que foi o dia marcado pelo início da Crise de 1929. Dessa forma, os comerciantes procuraram associar o termo a algo positivo e, como o dia depois do Thanksgiving é muito lucrativo, eles decidiram utilizar “Black Friday” para refletir o sucesso de vendas. Assim, vermelho (red) normalmente significa perda e preto (black) significaria ganho.

A Black Friday será o domínio da mineração de texto e análise de sentimento deste trabalho de conclusão de curso, onde serão extraídos conhecimentos a partir da análise das opiniões das pessoas postadas no Twitter.

3 Aplicação no domínio

Este capítulo descreverá a aplicação do processo de mineração de texto e análise de sentimentos no domínio da Black Friday. Para isso, será seguido o processo de mineração de texto proposto por Aranha (2007) e Konchady (2006) descrito na seção 2.1.1. Como o processo definido é genérico, o mesmo foi ligeiramente adaptado neste capítulo para atender as necessidades do domínio estudado neste trabalho.

3.1 Coleta

A coleta dos *tweets* para a mineração de texto e análise de sentimentos ocorreu através da API Tweepy 3.5 do Python, utilizando o módulo de streaming, ou seja, capturando os comentários à medida em que eles eram postados no momento da coleta. Os *tweets* foram coletados no dia 27 de novembro de 2015 no período da tarde por cerca de 10 minutos, totalizando 30.498 *tweets*. Os termos utilizados para coletar os comentários sobre a Black Friday foram “blackfriday” e “black friday”, pois, segundo a documentação oficial do Twitter (2015), um *tweet* vai corresponder aos termos de busca quando os mesmos estiverem presentes no texto independente da ordem em que foram colocados no filtro ou se estão em maiúsculo ou minúsculo. Dessa forma, nos resultados coletados apareceram comentários com os termos “#blackfriday”, “#BlackFriday”, “Black Friday” e suas variações.

Cada *tweet* foi coletado no formato JSON e possui atributos como data e hora da criação, id, texto, idioma, nome do usuário, localidade, entre outros, como mostra a Figura 7. Porém, nem todos esses atributos serão necessários para a realização da mineração de texto e análise de sentimentos. A etapa de pré-processamento será responsável por justamente selecionar os atributos que serão relevantes para o processamento dos textos na fase seguinte, bem como segmentá-los e remover as stop words (palavras consideradas irrelevantes para a realização de uma análise de texto, por exemplo) dos mesmos.

```

{
  "created_at": "Fri Nov 27 16:16:26 +0000 2015",
  "id": 670274995935969284,
  "id_str": "670274995935969284",
  "text": "Nothing better than Black Friday",
  "source": "<a href='\"http://twitter.com/download/iphone\"' rel='\"nofollow\"'>Twitter for iPhone</a>",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 566543429,
    "id_str": "566543429",
    "name": "eden",
    "screen_name": "edenholmes",
    "location": "Hertfordshire",
    "url": null,
    "description": "ig: e11jh | you win some, you lose some",
    "protected": false,
    "verified": false,
    "followers_count": 361,
    "friends_count": 361,
    "listed_count": 0,
    "favourites_count": 17543,
    "statuses_count": 3189,
    "created_at": "Sun Apr 29 18:19:12 +0000 2012",
    "utc_offset": -28800,
    "time_zone": "Pacific Time (US & Canada)",
    "geo_enabled": true,
    "lang": "en",
    "contributors_enabled": false,
  }
}

```

Figura 7 – Parte de um arquivo json com alguns campos de um “tweet”

Fonte: Elaborado pelo ator

3.2 Pré-processamento

Como mencionado anteriormente, há atributos em cada *tweet* do arquivo JSON que não serão importantes na etapa de mineração. Devido à dificuldade em achar um corpus (conjunto de textos pré-classificados) em português, que será útil na etapa de treinamento dos *tweets*, e o fato de os textos coletados serem em sua maior parte em inglês, foi decidido que a mineração e a análise de sentimentos seriam feitos com base nos *tweets* em língua inglesa. Assim, do conjunto coletado, primeiramente, foram extraídos apenas os campos referentes ao comentário (“text”) e idioma (“lang”) e depois foram filtrados apenas os comentários com o termo “en” (English), sobrando do arquivo apenas o campo “text”. Dos 30.498 *tweets* coletados, 29.025 são em inglês.

3.2.1 Remoção de tweets repetidos

Devido à enorme quantidade de *retweets* dados pelos usuários no Twitter, muitos *tweets* encontram-se repetidos no conjunto coletado pela API. Não faz sentido mantê-los no conjunto, uma vez que os mesmos podem alterar os resultados gerados pela mineração e

análise de sentimentos. Assim, todos os *tweets* repetidos foram removidos. Dos 29.025 que restaram da filtragem anterior, sobraram, agora, 19.158 *tweets*.

3.2.2 Transformação de tweets em formato Unicode

Foi observado que alguns *tweets* possuíam símbolos em formato Unicode (Figura 8) e foi necessário adequá-los para que pudessem ser processados nas etapas seguintes. Após isso, caracteres como “u2019” foram transformados em apóstrofe, por exemplo. Além disso, muitos *tweets* possuíam “emojis” (emoticons) e pontuações representadas em Unicode que não serão relevantes na mineração de texto e nem na análise de sentimentos. Dessa forma, esses caracteres foram removidos dos textos por meio de expressões regulares.

```
Exhausted from #BlackFriday shopping. \ude27  
Pray for the Black Friday workers \ude1e \ude10 \ude14 \ude29 #LetUsBeSafe #LetPeopleBeNice  
RT @VSPINK: It\u2019s happening now, like right NOW! #ShopSoHard #BlackFriday
```

Figura 8 – Tweets com caracteres em Unicode

Fonte: Elaborado pelo autor

3.2.3 Transformação de tweets para letras minúsculas

Naturalmente, os textos possuíam muitas palavras com letras maiúsculas, que poderiam dificultar o processamento mais tarde, uma vez que palavras como “Good”, por exemplo, podem ser diferenciadas de “good” no momento da classificação do texto. Assim, todas as palavras de cada *tweet* foram transformadas para letras minúsculas.

3.2.4 Transformação de acrônimos e abreviações de internet

Também é comum em comentários de redes sociais os usuários utilizarem abreviações e acrônimos para se referirem a determinadas palavras. Acrônimos como “Gr8”, “L8”, “B4”, por exemplo, significam, respectivamente, “great”, “late” e “before”. Já as abreviações podem ser menos intuitivas de serem entendidas como “IMO”, que quer dizer “in my opinion”. Dessa forma, foram identificadas nos “tweets” algumas abreviações e acrônimos, de acordo com as listas elaboradas por Lee (2015) e Washenko (2015). Após a identificação, cada abreviação e acrônimo foram transformados para seus respectivos significados. Porém, os termos “HMU”, “HT”, “ICYMI”, “IRL”, “OOTD”, “RT”, “r”, “u” e “yo” foram eliminados dos *tweets*, pois seus significados foram considerados irrelevantes para a classificação do texto. As principais

abreviações e acrônimos encontrados nos comentários estão apresentados na Tabela 1. Essa transformação foi necessária, pois as etapas de remoção de stop words e análise de sentimentos não conseguiriam identificar esses termos e, conseqüentemente, os resultados obtidos não refletiriam a realidade.

Tabela 1 - Principais acrônimos e abreviações de internet encontrados nos “tweets”

Acrônimo/Abreviação	Significado
R	Are
Bc	Because
b/c	Because
b4	Before
BFF	Best friends forever
BTW	By the way
DM	Direct message
Fb	Facebook
FTW	For the win
FYI	For your information
gr8	Great
HMU	Hit me up
HT	Hat tip
ICYMI	In case you missed it
IDC	I don't care
IDK	I don't know
Ig	Instagram
ILY	I love you
IMO	In my opinion
Irl	In real life
JK	Just kidding
LMAO	Laughing my ass off
LOL	Laughing out loud
NVM	Never mind
OMG	Oh my God
OOTD	Outfit of the day

Ppl	People
RT	Retweet
SMH	Shaking my head
TBH	To be honest
TGIF	Thank God it's Friday
Thx	Thanks
til	Until
U	You
w/	With
w/o	Without
WTF	What the fuck
Y	Why
Yo	Your

Fonte: Elaborado pelo autor

Agora, os textos já estão prontos para passarem pelas duas técnicas de PLN que serão utilizadas neste trabalho: tokenização e remoção de stop words.

3.2.5 Tokenização

Após a pré-adequação dos *tweets* ao transformar as palavras em formato Unicode, letras maiúsculas em minúsculas e substituir alguns acrônimos e abreviações por seus significados, os textos passarão pelo processo de tokenização, ou seja, os *tweets* serão segmentados pelos espaços em branco entre as palavras. Para isso, será seguido o processo proposto por Konchady (2006) descrito na seção 2.1.1.2 deste trabalho com algumas alterações. A etapa de identificação de tokens multi-vocabulares será realizada mais tarde utilizando a técnica conhecida como Bigramas. Será utilizada a API NLTK 3.0.4, bastante conhecida para o processamento de linguagem natural, na linguagem de programação Python para a realização do processo de tokenização.

3.2.5.1 Geração simples de tokens

Nessa etapa, foram gerados os tokens da maneira mais simples, ou seja, cada token é composto por cada palavra, pontuação ou caractere especial contido em cada *tweet*. A Figura 9 apresenta um exemplo de um *tweet* e seus respectivos tokens após a tokenização.

Tweet:

"there aren't really any good deals for black friday. just propaganda to get people to buy stuff."

Tokens:

['there', 'are', 'n't', 'really', 'any', 'good', 'deals', 'for', 'black', 'friday', '.', 'just', 'propaganda', 'to', 'get', 'people', 'to', 'buy', 'stuff', '.']

Figura 9 – Exemplo de um “tweet” após passar pelo processo de tokenização

Fonte: Elaborado pelo autor

Esse exemplo ilustra um dos problemas de uma tokenização simples que é a separação de “aren’t” em dois tokens “are” e “n’t”. O ideal seria que esse termo estivesse no mesmo token. Dessa forma, todos os termos que se encontravam em sua forma contraída foram realocados no mesmo token. A Tabela 2 mostra uma lista com as principais formas contraídas.

Tabela 2 – Termos em inglês na forma contraída

Termos na forma contraída				
isn't	won't	can't	I'm	I/he/she/it/you/we'll
wasn't	couldn't	hadn't	you/we're	could've
weren't	wouldn't	don't	he/she/it's	would've
haven't	mightn't	doesn't	I/you/we've	must've
hasn't	mustn't	didn't	I/he/she/it/you/we'd	might've

Fonte: Elaborado pelo autor

3.2.5.2 Identificação de Abreviações

Os *tweets* também podem conter abreviações separadas por ponto que serão divididas durante o processo de tokenização. Por exemplo, a abreviação “U.S.”, após a tokenização, seria separada da seguinte forma:

[U][.][S][.]

Assim, com a ajuda de expressão regular, foram identificadas as abreviações presentes nos *tweets* e colocadas no mesmo token. A Tabela 3 apresenta uma lista com as principais abreviações encontradas.

Tabela 3 - Lista com as principais abreviações dos “tweets”

Abreviações
U.S.
A.K.A.
A.M.
P.M.
P.S.
E.G.
R.I.P.
T.V.

Fonte: Elaborado pelo autor

3.2.5.3 Identificação de Palavras Combinadas

Algumas palavras podem estar combinadas com outras através dos caracteres especiais “&” e “-“. Em alguns tweets, por exemplo, a loja “H&M” é citada. Ao passar pela tokenização, os caracteres são separados e cada um ocupa um token diferente. Assim, utilizando expressão regular, foi identificada cada palavra combinada e estas passaram a ocupar o mesmo token.

3.2.5.4 Identificação de Símbolos de Internet e Números

Nessa etapa são identificados endereços de e-mail, links, hashtags, menções (@nome_do_usuario) dos *tweets* e todas as representações de números, incluindo medidas e valores. Como nas etapas anteriores, os tokens identificados com essas características foram reagrupados em um único token. Após a identificação, esses tokens foram eliminados do conjunto de cada *tweet*, pois não são termos relevantes para a classificação dos textos. As palavras que faziam parte das hashtags foram mantidas e somente o símbolo “#” foi removido.

Os *tweets* já estão prontos para a próxima etapa que é a de remoção de stopwords, pontuações e caracteres especiais.

3.2.6 Remoção de stopwords, pontuações e caracteres especiais

Essa é a última etapa de pré-processamento deste trabalho. A remoção de stopwords busca reduzir o número de palavras deixando apenas o conteúdo relevante para ser processado na etapa de mineração de texto e análise de sentimento. No contexto de textos da língua inglesa, palavras como “of”, “at”, “by” e “for”, por exemplo, só servem para ligar uma palavra a outra e completar seu significado. Neste trabalho, elas não serão importantes, pois não carregam nenhum sentimento e não farão diferença na hora da mineração e classificação dos *tweets*.

Existem várias listas de stopwords pela Internet, comumente chamadas de stoplists. A escolhida para este trabalho foi a stoplist da API NLTK 3.0.4, pois apresenta palavras suficientes para a realização da tarefa no domínio analisado. A Tabela 4 mostra a stoplist completa utilizada.

Tabela 4 – Lista de stopwords da API NLTK 3.0.4

Stoplist									
I	he	Their	are	doing	At	below	then	more	too
Me	him	Theirs	was	a	By	To	once	most	very
My	his	Themselves	were	an	For	from	here	other	s
Myself	himself	What	be	the	with	Up	there	some	t
We	she	Which	been	and	about	down	when	such	can
Our	her	Who	being	but	against	In	where	no	will
Ours	hers	Whom	have	if	between	out	why	nor	just
Ourselves	herself	This	has	or	into	On	how	not	don
You	it	That	had	because	through	Off	all	only	should
Your	its	These	Having	as	during	over	any	own	now

			g						
Yours	itself	Those	do	until	before	under	both	same	
Yourself	they	Am	does	while	after	again	each	so	
Yourself	them	Is	did	of	above	Further	few	than	

Fonte: Elaborado pelo autor

Após a remoção das stopwords, ainda falta remover as pontuações e os caracteres especiais presentes no texto. Ao fazer a tokenização, as pontuações e os caracteres especiais são colocados em tokens individuais, ou seja, dessa forma fica ainda mais fácil a identificação dos mesmos para a remoção. Para isso, foi utilizada a constante do tipo String presente no Python que possui uma lista com todas as pontuações e caracteres especiais, apresentada na Tabela 5.

Tabela 5 – Lista de pontuações e caracteres especiais

Lista de pontuações e caracteres especiais			
'	(:]
!)	;	^
"	*	<	_
#	+	=	`
\$,	>	{
%	-	?	
&	.	@	}
\	/	[~

Fonte: Elaborado pelo autor

Finalmente, os tweets estão prontos para a etapa de mineração de texto e análise de sentimento.

3.3 Mineração de texto e análise de sentimentos

Para realizar a etapa de mineração e análise de sentimento foi escolhido o algoritmo de Naïve Bayes, como descrito no capítulo anterior. Este algoritmo, implementado pela API

NLTK 3.0.4, foi utilizado para a classificação dos *tweets* coletados em positivo ou negativo, de acordo com o sentimento identificado pelo mesmo. A técnica de classificação escolhida para este trabalho foi a abordagem com Aprendizado de Máquina Supervisionado, ou seja, foi adotado um corpus com 2.000 reviews de filmes pré-classificadas (1.000 positivas e 1.000 negativas). Destas 2.000 reviews, 1.500 (750 positivas e 750 negativas) foram utilizadas para o treinamento com o classificador Naïve Bayes e as outras 500 (250 positivas e 250 negativas) foram usadas como teste para verificar a acurácia do classificador.

Foi escolhido o corpus de treinamento com reviews de filmes, pois não foi encontrado nenhum corpus que representasse ou fosse próximo do domínio da Black Friday. Além disso, dentre os corpus observados para a realização do treinamento, este foi o que possuía os dados mais completos e as classificações mais coerentes. Alguns corpus de treinamento classificados em três categorias (positivo, negativo e neutro), a princípio, também foram levados em conta, porém, foram logo descartados, pois seus dados também estavam incompletos e incoerentes. Nesse cenário, mesmo com os riscos de utilizar um corpus num domínio diferente do tratado no trabalho, foi decidido seguir adiante com o mesmo.

3.3.1 Treinamento do corpus

O treinamento foi baseado nos artigos publicados por Perkins (2010) em seu blog “StreamHacker”. Antes de realizar o treinamento do corpus foi necessária a realização de extração de *features*. Para isso, foi utilizado um modelo “Bag of words” simples, onde cada palavra representa uma *feature* e seu valor é representado por True, ou seja, cada *feature* representa a existência de uma palavra no corpus (PERKINS, 2010).

Após a extração de *features*, foi realizado o treinamento do corpus, onde foram utilizados os tokens com os *features* no formato {word: True} e os rótulos “pos” ou “neg”. Logo em seguida, o conjunto de reviews para teste foi classificado utilizando o conjunto treinado. Foi constatada uma acurácia de aproximadamente 73%. Além disso, utilizando a precisão e o recall para validar a efetividade dessa classificação, foi observado que quase toda review positiva foi corretamente identificada como positiva, com 98% de recall, ou seja, poucos falsos negativos. Porém, apenas 65% das reviews que foram classificadas como positivas estão corretamente classificadas, o que indica que podem ter 35% de falsos positivos para a label “pos” de acordo com a precisão. Para as reviews negativas, 96% podem ter sido identificadas corretamente, com poucos falsos positivos, mas várias reviews que são negativas foram classificadas incorretamente, com 52% de falsos negativos para a label “neg”.

Uma possível causa seria a existência de stopwords, que poderiam estar causando algum ruído nos corpus de treinamento e de teste. Dessa forma, como teste, foram retiradas todas as stopwords das reviews. Agora, a acurácia caiu 0,2% e tanto a precisão quanto o recall também caíram. Ou seja, a remoção de stopwords, nesse caso, não melhorou nenhuma das métricas.

Outra possível causa desses falsos negativos e positivos terem sido altos pode ser a existência de palavras positivas em reviews negativas precedidas de palavras negativas como, por exemplo, “not good”. Assim, como foi utilizada a técnica de “Bag of words” em que cada palavra é independente, o classificador não foi capaz de perceber essas palavras combinadas. Dessa forma, com o intuito de aumentar a precisão, o treinamento foi testado, também, utilizando a técnica conhecida como “Bigramas” (“Bigrams”), que produzirá palavras combinadas baseadas em suas frequências nos textos.

A eliminação de *features* também pode melhorar a classificação, uma vez que se o modelo tem milhares de *features*, a probabilidade dele conter informação irrelevante é enorme, além de diminuir a performance. Assim, foi utilizado no treinamento apenas as 10.000 palavras mais relevantes que aparecem nos textos, já utilizando, também, os 200 melhores “bigramas”.

Após o treinamento, a acurácia aumentou de 73% para 93%, enquanto a precisão da label “pos” aumentou para 89% e o recall permaneceu o mesmo. Já o recall da label “neg” aumentou para 88% e a precisão aumentou para 98%.

Para conseguir um índice ainda maior, foi utilizada a técnica de Ganho de Informação para cada palavra. Em classificação de texto, Ganho de Informação mede o quão comum é uma *feature* em uma classe específica em relação a todas as outras classes. Por exemplo, a presença da palavra “magnificent” numa review de filme é uma forte indicação de que a review é positiva, o que faz essa palavra ser considerada relevante. Após a adição dessa técnica, a acurácia diminuiu 1%. Porém, a precisão e o recall da label “pos” aumentaram para, respectivamente, 91% e 93%, enquanto a precisão e o recall da label “neg” aumentaram para 93% e 92%, respectivamente. Isso significa que a utilização de “Bigramas” e a seleção dos *features* mais relevantes utilizando a técnica de Ganho de Informação foram boas alternativas para melhorar a classificação, nesse caso. A Tabela 6 mostra um resumo das medidas para cada tipo de abordagem utilizada.

Tabela 6 – Resumo das medidas utilizadas para validar o classificador

Abordagem	Acurácia	Precisão “pos”	Precisão “neg”	Recall “pos”	Recall “neg”
Bag of words	73%	65%	52%	98%	96%
Bag of words e remoção de stopwords	72,6%	65%	96%	98%	48%
Bigramas e seleção de features	93%	89%	98%	98%	88%
Bigramas, seleção de features e ganho de informação	92%	91%	93%	93%	91%

Fonte: Elaborado pelo autor

3.3.2 Análise de sentimento

Após o treinamento do corpus e a verificação da qualidade do classificador através da acurácia, precisão e recall, os mais de 19.000 *tweets* coletados sobre a “Black Friday” foram classificados em positivo ou negativo pelo classificador. Em seguida, 20 *tweets* foram escolhidos aleatoriamente, classificados manualmente e comparados com os sentimentos atribuídos pelo classificador. A Tabela 7 apresenta a comparação.

Tabela 7 – Comparação entre os sentimentos classificados automaticamente e manualmente

Tweets	Manualmente	Classificador
“i don't do black friday. went once a couple years ago..had an anxiety attack and i haven't gone since. and don't plan to”	neg	neg
“@happenings9ja: #blackfriday: insane, awesome, unbelievable!!! let me tell you how it all started!”	pos	pos
“could not think of anything worse than going shopping on black friday”	neg	neg

“black friday shopping going great”	pos	neg
“i officially hate black friday!”	neg	pos
“the black friday sales sucked to be honest lol”	neg	neg
“black friday is incredibly lame in germany”	neg	neg
“i love black friday!!!”	pos	neg
“i love black friday i bought so much shit today”	pos	pos
“don't get me wrong i love a sale , but black friday is not for me”	neg	pos
“i hate black friday. i would rather buy full price items than go shopping on this day.”	neg	pos
“@empressilversky i get to work this afternoon. i hate black friday... i would much rather binge watch tv shows lol”	neg	neg
“most successful black friday ever”	pos	pos
“best buy canada site malfunctions on black friday, frustrating shoppers”	neg	neg
“@bridgeyrozz: black friday was not even as bad as i thought it would be”	pos	neg
“man i hate black friday so bad”	neg	neg
“i'm def the only person alive who doesn't care about black friday this year lol”	neg	neg
“these black friday deals are too good to miss! #confidence #hairgrowthsolutions #blackfriday”	pos	neg
“one of america's greatest humiliations, #blackfriday.”	neg	pos
“i love #blackfriday cause while everyone is shopping the gym is empty! #fitfam #havefunfightingthecrowd”	pos	neg

Fonte: Elaborado pelo autor

Como pode ser observado, dos 20 tweets classificados automaticamente, 11 foram classificados corretamente e 9 incorretamente, ou seja, o classificador conseguiu acertar a polaridade de mais da metade dos tweets selecionados. Porém, o classificador errou a

polaridade do *tweet* “i love black friday!!!”, uma frase simples e que deveria ter sido classificada como positiva. Da mesma forma, o *tweet* “i officially hate black friday!”, que deveria ter sido classificado como negativo, foi classificado como positivo, ou seja, ocorreu um erro grave de classificação que teve ser corrigido. Para isso, foi necessário retornar à fase de treinamento para que pudessem ser feitos ajustes.

A alternativa escolhida para melhorar a classificação dos *tweets* foi a diminuição das *features* e dos bigramas da etapa de treinamento. Assim, foram selecionadas apenas as 1250 palavras mais relevantes e os 50 melhores bigramas. Após realizar o treinamento novamente, a acurácia diminuiu de 92% para 88%, a precisão da label “pos” também foi de 88% e o recall ficou em 87%. Já a precisão e o recall da label “neg” foram para 88%. Os números de acurácia, precisão e recall diminuíram em relação ao resultado adotado anteriormente, mas pelo menos dessa vez os resultados da classificação fazem mais sentido. Agora, dos 20 *tweets* observados, apenas 7 foram classificados incorretamente. A Tabela 8 apresenta esses resultados.

Tabela 8 – Comportamento dos tweets classificados após novo treinamento

Tweets	Manualmente	Classificador
“i don't do black friday. went once a couple years ago..had an anxiety attack and i haven't gone since. and don't plan to”	neg	neg
“@happenings9ja: #blackfriday: insane, awesome, unbelievable!!! let me tell you how it all started!”	pos	neg
“could not think of anything worse than going shopping on black friday”	neg	neg
“black friday shopping going great”	pos	pos
“i officially hate black friday!”	neg	neg
“the black friday sales sucked to be honest lol”	neg	neg
“black friday is incredibly lame in germany”	neg	neg
“i love black friday!!!”	pos	pos
“i love black friday i bought so much shit today”	pos	pos
“don't get me wrong i love a sale , but black friday is not for me”	neg	pos
“i hate black friday. i would rather buy full price	neg	neg

items than go shopping on this day.”		
“@empresilverky i get to work this afternoon. i hate black friday... i would much rather binge watch tv shows lol”	neg	neg
“most successful black friday ever”	pos	pos
“best buy canada site malfunctions on black friday, frustrating shoppers”	neg	pos
“@bridgeyrozz: black friday was not even as bad as i thought it would be”	pos	neg
“man i hate black friday so bad”	neg	neg
“i'm def the only person alive who doesn't care about black friday this year lol”	neg	pos
“these black friday deals are too good to miss! #confidence #hairgrowthsolutions #blackfriday”	pos	neg
“one of america's greatest humiliations, #blackfriday.”	neg	pos
“i love #blackfriday cause while everyone is shopping the gym is empty! #fitfam #havefunfightingthecrowd”	pos	pos

Fonte: Elaborado pelo autor

Como esperado, o classificador não é perfeito, mas também possui algumas limitações que serão discutidas mais adiante. Obviamente, não é possível validar toda a análise de sentimento realizada com base só em 20 tweets, uma vez que foram classificados mais de 19.000. No próximo capítulo, serão analisados os resultados da análise de sentimento e de outras informações provenientes dos tweets sobre a Black Friday.

4 Análise dos resultados

A última etapa do processo de mineração de textos proposto por Aranha (2007) é a de análise de resultados. Neste trabalho, essa etapa será discutida e apresentada através de textos, tabelas, figuras e gráficos.

Como qualquer tarefa no campo da mineração de texto, processamento de linguagem natural e análise de sentimentos, os resultados sempre devem ser encarados apenas como um indicador e não como verdade absoluta. O corpus de treinamento é limitado e isso quer dizer que há palavras e expressões que podem não ter sido identificadas na etapa de classificação dos *tweets*. Além disso, a linguagem falada na Internet possui gírias, ironias e sarcasmo que não são reconhecidos em modelos simples de análise de texto. Os *tweets* também possuem muitos erros de digitação que dificultam a classificação, pois o classificador também não reconhece essas palavras.

No caso deste trabalho, os *tweets* foram classificados em positivos e negativos. Porém, nem todos os *tweets* expressam opiniões positivas ou negativas. Alguns deles apresentam apenas fatos ou comentários ambíguos que não cabem em nenhuma das duas opções. Como o conjunto de treinamento escolhido classificava os textos apenas em positivos ou negativos, os *tweets* também só foram classificados com esses dois rótulos.

Dos 19.158 *tweets* utilizados, 9.637 foram classificados como positivos e 9.521 como negativos, ou seja, aproximadamente 51% dos *tweets* classificados são positivos, enquanto 49% são negativos (Figura 10). De maneira geral, isso significa que, segundo o resultado da análise de sentimento, apesar dos resultados serem bem próximos, as pessoas fizeram mais comentários positivos que negativos sobre a Black Friday no Twitter.

Uma análise realizada pelo site AdWeek (2014) em relação à Black Friday de 2014 concluiu que 18% dos comentários no Twitter e Facebook durante todo o dia 28 de novembro de 2014, data da Black Friday, foram positivos e 14% negativos. Porém, a maior parte dos comentários foram classificados como neutros, com 68%. Comparando com os resultados deste trabalho, isso pode significar que muitos *tweets* classificados como positivos ou negativos foram classificados incorretamente, uma vez que a presença de comentários neutros na pesquisa do site é maior que a presença de comentários positivos e negativos. Apesar disso, tem que ser levado em conta que o site realizou a pesquisa em duas redes sociais e por 24 horas, ou seja, a quantidade de comentários e, conseqüentemente, de sentimentos observados é bem maior.

Já a análise realizada pelo site DataRank (2015) em relação à Black Friday de 2015 concluiu que 60% dos comentários coletados no Twitter entre os dias 26 e 28 de novembro (a Black Friday em 2015 aconteceu dia 27) são positivos, enquanto 40% são negativos. Tirando o fato da quantidade de tweets coletados pelo site (9.084.626) e a quantidade de dias em que a coleta foi realizada (3 dias) e comparando com os resultados deste trabalho, pode-se concluir que, apesar de ambos os resultados possuírem uma quantidade de classificações incorretas devido a ausência de classificação de comentários neutros, ambos chegaram a mesma conclusão.

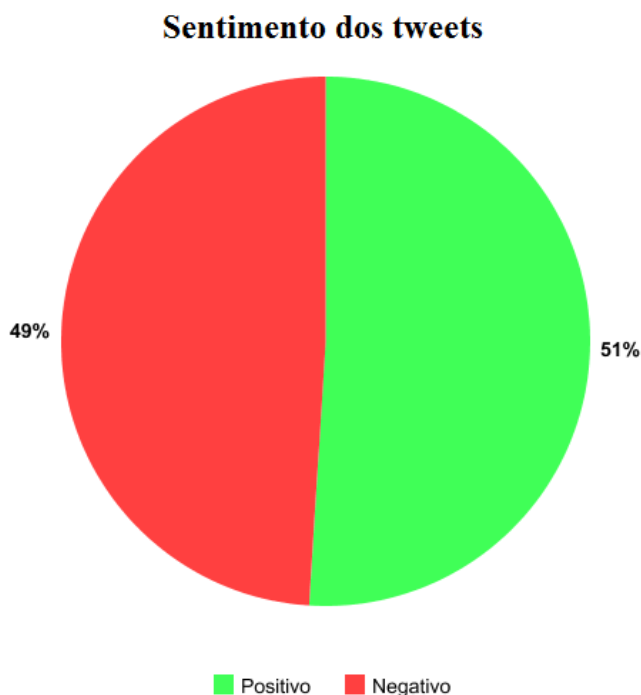


Figura 10 – Gráfico com o sentimento dos tweets sobre a Black Friday

Fonte: Elaborado pelo autor

Analisando as palavras mais frequentes dos *tweets* o resultado foi que “shopping” (compras), “sale” (venda) e “deals”(promoções) foram as palavras que mais foram mencionadas pelos usuários do Twitter nos dados coletados, o que compreensível, já que os *tweets* tratam-se de comentários sobre a Black Friday. A Figura 11 apresenta uma nuvem de palavras com as 300 palavras mais frequentes nos *tweets*. Foram retiradas das palavras mais frequentes os termos “black”, “friday” e “blackfriday”, pois foram termos utilizados para a coleta dos comentários e aparecem em todos os *tweets*.

Nota-se que muitas outras palavras que aparecem em destaque na nuvem também são termos relacionados à compra e venda como “discount” (desconto), “buy” (comprar),

de acordo com a AdWeek, as marcas mais comentadas foram Amazon, Walmart e Best Buy. A Tabela 9 apresenta as oito marcas e lojas que foram mais mencionadas pelos usuários nos tweets coletados. É importante ressaltar que as “menções” dizem respeito à frequência que as palavras aparecem nos comentários e não às contas do Twitter que tiveram mais menções. Como pode ser observado, o número da frequência de cada marca mostrado na tabela é baixo em relação à quantidade de *tweets*. Uma provável explicação seria a de que muitos usuários comentaram sobre uma determinada marca indiretamente, não mencionando a mesma num comentário ou se referiram no *tweet* a um produto específico da marca, mas não citaram o nome da mesma.

Tabela 9 – Marcas mais comentadas nos tweets

Marca	Frequência
Amazon	296
Target	119
Walmart	98
Apple	54
Best Buy	44
Samsung	20
Kohl's	17
Macy's	12

Fonte: Elaborado pelo autor

Entre os *tweets* classificados como positivos, a Amazon permanece em primeiro lugar como a marca mais comentada, seguida pela Target e Walmart. As marcas com menos *tweets* positivos são Macy's e Kohl's. As três primeiras tem uma forte atuação no Twitter e também são conhecidas pelas melhores promoções durante a Black Friday. Macy's e Kohl's, as duas últimas colocadas, não possuem escala global como a Amazon, Target e Walmart, atuando apenas nos Estados Unidos, Porto Rico e Guam, ou seja, muitos comentários que mencionavam as lojas Macy's e Kohl's podem ter se restringido apenas a esses países. A Tabela 10 mostra a frequência de menções das marcas em tweets classificados como positivos.

Tabela 10 – Marcas mais comentadas nos tweets classificados como positivos

Marca	Frequência
Amazon	136
Target	91
Walmart	45
Apple	28
Best Buy	23
Samsung	11
Macy's	9
Kohl's	7

Fonte: Elaborado pelo autor

As marcas mais comentadas em *tweets* negativos não sofreram muitas mudanças em relação aos *tweets* positivos. Amazon, Walmart e Target permanecem nos primeiros lugares. Porém, quase a metade dos *tweets* que mencionaram a Amazon e mais da metade dos comentários sobre o Walmart são negativos. Isso pode ter acontecido, pois muitas pessoas encontraram problemas ao acessar o site dessas lojas, devido à grande quantidade de acesso ao mesmo tempo em, com isso, foram até o Twitter para reclamarem sobre o ocorrido. Já as marcas com menos *tweets* negativos são a Samsung e a Macy's, que não possuem uma grande quantidade de menções no total. A Tabela 11 ilustra as marcas mais comentadas em *tweets* negativos.

Tabela 11 – Marcas mais comentadas nos tweets classificados como negativos

Marca	Frequência
Amazon	130
Walmart	53
Target	28
Apple	26
Best Buy	21
Kohl's	11
Samsung	9
Macy's	5

Fonte: Elaborado pelo autor

A análise realizada neste trabalho representa apenas os resultados obtidos durante a coleta dos tweets e não o de toda a Black Friday 2015.

5 Conclusão

Atualmente, a mineração de texto e análise de sentimento tornaram-se tarefas imprescindíveis para as empresas que querem conhecer o que os seus clientes estão falando da sua marca nas redes sociais, ao mesmo tempo em que sejam capazes de extrair informações que os possibilitem traçar estratégias competitivas que auxiliem em seus negócios.

Este trabalho apresentou um processo de mineração de texto e análise de sentimento ao analisar os comentários coletados do Twitter sobre a “Black Friday”. Os comentários foram coletados no dia do evento, 27 de novembro de 2015, por cerca de 10 minutos através da API de Streaming Tweepy. Mais de 30.000 *tweets* foram coletados, mas após os mesmos passarem por processos de filtragem e pré-processamento, restaram apenas 19.158 *tweets* para análise. Os *tweets*, então, foram classificados em positivos ou negativos utilizando o algoritmo de Naïve Bayes da API NLTK para análise de sentimento. Dessa forma, os resultados foram analisados e comparados com dados de outras pesquisas.

Apesar da ótima acurácia, precisão e recall promovidos pela etapa de treinamento, muitos *tweets* foram classificados incorretamente. Um dos motivos para essa má classificação pode ter sido originado do corpus de treinamento. O corpus escolhido era sobre reviews de filmes e provavelmente não reconhece todas as palavras do vocabulário utilizado pelos usuários do Twitter ao comentarem sobre a Black Friday, mesmo que ainda seja capaz de identificar opiniões positivas e negativas. Além disso, o corpus foi classificado levando em conta apenas duas categorias, positivo e negativo, o que obriga o classificador a categorizar os *tweets* coletados apenas nessas duas categorias. Porém, foi observado que nem sempre os comentários dos usuários possuíam opiniões positivas ou negativas e que, em alguns casos, apresentavam apenas fatos. Dessa forma, foi concluído que a utilização de apenas duas categorias para analisar o sentimento dos *tweets* sobre a Black Friday não foi a abordagem mais adequada. A existência de gírias, sarcasmo, ironia e erros de digitação também dificultaram a classificação, uma vez que o classificador não era capaz de entender ambiguidades e palavras que não existissem em seu “dicionário”.

Apesar das dificuldades, os resultados finais obtidos, em comparação com outras pesquisas, estão, em pequena escala, bem próximos. Muitos *tweets* foram classificados corretamente e contribuíram para a geração do conhecimento do domínio analisado. Em contrapartida, o processo de mineração precisa ser melhorado para que os resultados sejam mais bem sucedidos.

O campo de mineração de texto e análise de sentimento ainda está em processo de evolução e, como comentado no capítulo anterior, os resultados devem ser tratados como indicadores e não como verdades absolutas. Assim, de modo geral, o objetivo do trabalho, que era o de analisar a opinião dos usuários do Twitter sobre a Black Friday quanto ao sentimento atrelado aos comentários, foi atingido.

5.1 Trabalhos futuros

Para trabalhos futuros é sugerido a utilização de um corpus de treinamento no domínio analisado, para que ele tenha o máximo de conhecimento do domínio possível. Dessa forma, o classificador será capaz de identificar o sentimento com mais precisão. Além disso, seria ideal que o corpus tivesse textos pré-classificados em três categorias: positivo, negativo e neutro.

Os trabalhos seguintes também poderiam prestar mais atenção nas gírias, sarcasmos e ironias utilizados no Twitter, de modo a identificá-los e classificá-los de maneira correta. Análises de sentimento utilizando emoticons ou emojis, hoje em dia muito utilizados nas redes sociais para expressarem diversos sentimentos, também poderia ser uma abordagem interessante.

A utilização de outras APIs, como a AlchemyAPI da IBM, pode facilitar o trabalho de processamento de linguagem natural em trabalhos futuros de análise de texto.

Outra sugestão seria a realização da análise de sentimento em tempo real, de modo a apresentar a classificação dos sentimentos minutos depois que o *tweet* foi criado. Os resultados poderiam ser mostrados através de um site com gráficos, tabelas e outros artifícios que pudessem mostrar de maneira simples a polaridade dos tweets de determinado evento na hora em que ele acontece.

Referências Bibliográficas

HEARST, Marti. **What is Text Mining?** Disponível em: <<http://people.ischool.berkeley.edu/~hearst/text-mining.html>> Acesso em: 25 de outubro de 2015.

HOTH, A.; NURNBERGER, A.; PAASS, G. A Brief Survey of Text Mining. **LDV Forum - GLDV Journal for Computational Linguistics and Language Technology**, v. 20, n. 1, p. 19-62, 2005.

MINER, Gary, et al. **Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications**. 2012.

CONTRERAS, Roxana Jiménez. **Técnicas de Seleção de Características aplicadas a Modelos Neuro-Fuzzy Hierárquicos BSP**. 2002. 98 f. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2002.

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional**. 2007. 146 f. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Liddy, E.D. **Natural Language Processing**. In Encyclopedia of Library and Information Science. 2 ed. Nova York: Marcel Decker. 2001.

TRIM, Craig. **The Art of Tokenization**. Disponível em: <<https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization>> Acesso em: 15 de novembro de 2015.

KONCHADY, M. **Text Mining Application Programming**. Charles River Media. 1 ed. 2006.

CARRILHO JUNIOR, J. R. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. 96 f. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

GONÇALVES, João Carlos de Almeida Rodrigues. **Story Mining: Elicitação de Processos de Negócio a partir de Group Storytelling e Técnicas de Mineração de Texto**. Rio de Janeiro: Unirio, 2010. 175 f. Dissertação – Programa de Pós-Graduação em Informática da UNIRIO, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2010.

LIU, Bing. **Sentiment Analysis and Opinion Mining**. 2012.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**. v. 5, 1093-1113, 2014.

MURPHY, Kevin P. **Machine Learning: A Probabilistic Perspective**. 2012.

TOIT, Jurgens du. **The Bayes Classifier: building a tweet sentiment analysis tool**. Disponível em: <<http://cloudacademy.com/blog/naive-bayes-classifier/>> Acesso em: 22 de novembro de 2015.

NLTK. **Naive Bayes Module**. Disponível em: <<http://www.nltk.org/api/nltk.classify.html#module-nltk.classify.naivebayes>> Acesso em: 29 de novembro de 2015.

TWITTER. **Uso do Twitter/ Fatos sobre a empresa**. Disponível em: <<https://about.twitter.com/pt/company>> Acesso em: 24 de outubro de 2015.

ABOUT. **Why Is Black Friday Called Black Friday?** Disponível em: <http://useconomy.about.com/od/demand/f/Black_Friday_Name.htm> Acesso em: 29 de novembro de 2015.

G1. Brasileiros também terão os descontos da chamada 'Black Friday'. Disponível em: <<http://g1.globo.com/tecnologia/noticia/2010/11/brasileiros-tambem-terao-os-descontos-da-chamada-black-friday.html>> Acesso em: 29 de novembro de 2015.

TWITTER. Streaming API request parameters. Disponível em: <<https://dev.twitter.com/streaming/overview/request-parameters#track>> Acesso em: 5 de dezembro de 2015.

WASHENKO, Anna. The 75 Most Important Social Media Acronyms. Disponível em: <<http://sproutsocial.com/insights/social-media-acronyms/>> Acesso em: 16 de dezembro de 2015.

LEE, Kevan. The Definitive List of Social Media Acronyms and Abbreviations. Disponível em: <<https://blog.bufferapp.com/social-media-acronyms-abbreviations>> Acesso em: 16 de dezembro de 2015.

PERKINS, Jacob. Text classification for sentiment analysis – Naive Bayes Classifier. Disponível em: <<http://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/>> Acesso em: 27 de dezembro de 2015.

ADWEEK. Amazon Dominates Black Friday, Cyber Monday on Facebook, Twitter. Disponível em: <<http://www.adweek.com/socialtimes/engagor-amazon-black-friday-cyber-monday/439742>> Acesso em: 10 de janeiro de 2016.

DATARANK. The Highlights Of Black Friday 2015. Disponível em: <<https://blog.datarank.com/infographic-highlights-black-friday-2015/>> Acesso em: 10 de janeiro de 2015.