UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Igor Tannús Corrêa

Análise dos sentimentos expressos na rede social *Twitter* em relação aos filmes indicados ao *Oscar 2017*

Uberlândia, Brasil

2017

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Igor Tannús Corrêa

Análise dos sentimentos expressos na rede social *Twitter* em relação aos filmes indicados ao *Oscar 2017*

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Orientadora: Profa. Dra. Elaine Ribeiro de Faria Paiva

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2017

Igor Tannús Corrêa

Análise dos sentimentos expressos na rede social *Twitter* em relação aos filmes indicados ao *Oscar 2017*

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Uberlândia, Brasil, 08 de Dezembro de 2017:

Profa. Dra. Elaine Ribeiro de Faria Paiva
Orientadora

Prof. Dr. rer. nat. Daniel Duarte Abdala

Prof. Dr. Rodrigo Sanches Miani

Uberlândia, Brasil
2017

Agradecimentos

À minha orientadora, Elaine: obrigado por acreditar em mim como seu aluno. Graças à sua orientação, paciência e entusiasmo consegui finalizar este trabalho com êxito e como eu imaginava.

Aos meus pais: obrigado por sempre cuidarem de mim, por todos os sacrifícios que já fizeram para que eu e minhas irmãs tenhamos uma vida digna e por, desde sempre, terem me incentivado a estudar. Sem vocês, eu não teria alcançado nem metade das conquistas que eu tenho hoje.

Às minhas irmãs: obrigado por serem as melhores irmãs do mundo e por sempre tentarem arrancar um sorriso do meu rosto.

Aos meus outros familiares: obrigado por seu apoio e carinho.

Aos meus amigos: este último ano de graduação foi bem menos estressante do que eu esperava e muito mais divertido porque eu tenho vocês para festejarmos e falarmos bobagens. Obrigado por rirem das minhas piadas sem graça, por ouvirem os meus dramas e por estarem próximos quando preciso de vocês. Agradecimento especial a todos que me ouviram, sem reclamar, quando tudo que eu conversava era sobre *tweets*!

Aos meus professores da UFU: agradeço pelo seu tempo e dedicação, que foram essenciais à minha formação acadêmica. Graças a vocês, estou a poucos passos de conseguir o meu diploma em Sistemas de Informação pela Universidade Federal de Uberlândia.

A todos que transmitiram boas energias para mim e desejaram o meu bem, especialmente durante este ano que tem sido de suma importância para a minha formação acadêmica e carreira. Obrigado!

"Ciência é o que entendemos suficientemente bem para explicar a um computador.
Arte é tudo o mais que realizamos."

—Donald Knuth

Resumo

As redes sociais são espaços na *Internet* que possibilitam a criação e compartilhamento de conteúdo, sendo fontes importantes de informações e opiniões. A indústria cinematográfica é uma das que se beneficia das redes sociais, pois são utilizadas como meios de divulgação de filmes e também fontes de opiniões que podem influenciar próximos lançamentos e a forma como são divulgados.

Este trabalho visa realizar a análise dos sentimentos expressos pelos usuários da rede social *Twitter* em relação aos filmes indicados à categoria de Melhor Filme do *Oscar 2017*. Para isso, dados provenientes dessa rede foram coletados e, então, as etapas de uma tarefa em Análise de Sentimentos foram realizadas. Diferentes abordagens para classificação de textos foram estudadas e aplicadas a uma base de dados rotulada e, após avaliar os classificadores, o *Naive Bayes* multinomial foi escolhido para classificar a base completa.

A partir dos resultados, notou-se que esse método é interessante para realizar observações e análises sobre a base em questão, sendo bastante provável de se predizer qual filme seria o grande vencedor da premiação e quais estariam entre os menos prestigiados. Todavia, não foram encontradas associações matemáticas significativas entre o resultado do *Oscar 2017* e os resultados obtidos pela análise de sentimentos dos *tweets*. Também foi observado que os usuários do *Twitter* preferem usar a rede para publicar comentários positivos sobre filmes, ao invés de falar mal sobre os que não gostaram, e que premiações como o *Oscar* rendem uma grande quantidade de comentários no *Twitter*.

Palavras-chave: Análise de Sentimentos, *Twitter*, *Oscar 2017*, Mineração de Dados.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

As primeiras redes sociais surgiram no fim da década de 1990 e início dos anos 2000, se tornando tendências na *Web*. O grande crescimento da *Internet*, agregado à criação de novas tecnologias e políticas que facilitaram o seu acesso à população resultou no aumento da presença das redes sociais na vida das pessoas, o que mudou drasticamente a forma em que informações são transmitidas e acessadas pelos internautas (OLIVEIRA, 2013).

Os usuários das redes sociais agora não são mais apenas consumidores de informação, mas também geradores de conteúdo, já que lhes é oferecida a oportunidade de expressar as suas opiniões por meio de *blogs*, comentários, fórums e outras mídias sociais (FILHO, 2014). Essa disseminação de ideias acabou chamando a atenção de estudiosos e de empresas que procuram encontrar dados relevantes dentre a imensa quantidade de informação que é gerada a cada momento na rede (BOTHOS; APOSTOLOU; MENTZAS, 2010). Dessa forma, os usuários das redes sociais não se limitaram à interação com conhecidos, mas acabaram transformando-as em espaços na *Web* para se divulgar produtos e serviços, buscar e informar sobre acontecimentos políticos e catástrofes em tempo real, comentar sobre sua vida cotidiana, entre outros.

Em 2006, surgiu o *Twitter*[1] , uma rede social com uma premissa incomum – cada postagem, chamada de *tweet*, não deve exceder o limite de 140 caracteres. Devido à sua crescente popularidade, a rede gradativamente fez-se mais eficiente e acessível, tornando-se de suma importância, dando aos seus usuários o poder de criar e compartilhar ideias e informações instantaneamente, sem barreiras (TWITTER, 2017a).

Com mais de 313 milhões de usuários ativos mensalmente (TWITTER, 2017a), o *Twitter* é uma ferramenta fundamental em momentos de tensão política e manifestações. Durante a Primavera Árabe, em 2011, a população descontente usou o *Twitter* e outras mídias sociais para marcar encontros, divulgar protestos e mostrar ao mundo sua situação em tempo real (TAVARES, 2012). No Brasil, em 2013, a rede social também foi usada para divulgar protestos contra o aumento das tarifas de ônibus (COSTA, 2013), e em 2015 e 2016, para informar sobre as manifestações contra e a favor do *impeachment* da ex-presidente Dilma Rousseff (G1, 2016) (ROSSI, 2016).

Além disso, o *Twitter* também é um espaço para entretenimento, tendo muitos de seus usuários *tweetando*[2] sobre produtos, filmes, músicas, esportes, premiações, etc. Assim sendo, muitas companhias viram no *Twitter* a oportunidade de se tornarem mais próximas de seus clientes. No Brasil, a Netflix – líder no serviço de *streaming* de filmes e

[1] http://www.twitter.com

[2] O ato de publicar novos *tweets* em seu perfil na rede social *Twitter* é chamado de *tweetar*.

séries pela *Internet* –, Pontofrio – rede brasileira de varejo que vende móveis e eletrônicos – e o banco Santander são alguns exemplos que se destacam na rede social pela grande interação com seus seguidores. No entanto, a cada minuto, cerca de 350 mil *tweets* são publicados (STATS, 2017) e neles há muitas informações valiosas que possibilitam a essas organizações descobrir a opinião desses usuários com relação a seus produtos e serviços (FILHO, 2014).

Em razão da rapidez com que as informações são transmitidas na *Internet*, técnicas de Extração de Conhecimento são utilizadas para automatizar a busca e processamento de textos que, acompanhadas de técnicas de Análise de Sentimentos, tornam possível a descoberta do julgamento dos usuários com relação aos produtos, serviços e companhias. Consequentemente, as organizações são capazes de realizar melhorias e adotar práticas de acordo com a opinião de seu público-alvo.

Com fontes como o *Twitter*, que geram grandes volumes de dados a cada momento – também conhecidos como *big data* –, a Análise de Sentimentos tem se tornado cada vez mais relevante para essas companhias, assim como a busca por novas soluções que auxiliem na extração de conhecimento útil a partir dessas grandes bases de dados.

1.1 Justificativa

A indústria cinematográfica é bastante lucrativa no mundo inteiro, o que pode ser comprovado por meio das receitas de bilheterias, que totalizaram 36,4 bilhões de dólares em 2014 (MCHATTON, 2015). No Brasil, a indústria de TV e Filmes também é relevante, contribuindo com 19,8 bilhões de reais para a economia do país em 2013 e gerando 110 mil empregos em 2012 (BRAZIL, 2017). A situação também é satisfatória nos Estados Unidos, onde as bilheterias renderam 10,4 bilhões de dólares apenas em 2014 (MCHATTON, 2015).

Uma vez que redes sociais e entretenimento geralmente estão fortemente ligados, muitos usuários utilizam suas contas para expressar sua opinião, entusiasmo ou decepção com relação a filmes, seu elenco e produção. Toda essa euforia aumenta próximo da entrega dos *Oscars*, resultando na publicação de mais *tweets*, visto que essa é a premiação mais respeitada e comentada da indústria cinematográfica, além de ser responsável por impulsionar a divulgação dos filmes indicados e colaborar para ascenção da carreira dos profissionais e atores nomeados (WONG, 2013).

Assim sendo, é significativo saber se existe alguma correlação entre a opinião do público do *Twitter* e o resultado do *Oscar*. Essa análise pode ser feita usando técnicas de Análise de Sentimentos, uma área que, nos últimos anos, tem interessado vários pesquisadores e apresentado resultados promissores, como Tumasjan et al. (2010), que demonstraram como o *Twitter* pode ser um indicador em tempo real de sentimento político

enquanto Mandel et al. (2012) apresentaram essa rede social como uma fonte ao vivo de percepções públicas de um desastre natural.

A quantidade de trabalhos relacionados a Mineração de Opiniões e Análise de Sentimentos é abundante, por ser uma área que tem um objetivo excepcional, que é dar sentido à imensidão do *big data*. Cada vez mais esta tem sido uma área que gera retornos notáveis, o que inspirou pesquisadores como Benevenuto, Almeida e Silva (2011) e Cervi (2008) a explorarem os diferentes métodos e técnicas da área que podem ser aplicados às redes sociais.

Vários trabalhos têm sido desenvolvidos no intuito de correlacionar o sentimento positivo ou negativo dos *tweets* com relação a algum acontecimento. Em Filho (2014), o autor foi capaz de mapear o sentimento dos usuários do *Twitter* com relação aos jogos do Brasil na Copa do Mundo de 2014. Já um estudo publicado por Almeida (2012) expôs a grande incidência de *cyberbullying* realizada contra professores no *Twitter*.

Teixeira e Azevedo (2011) encontraram vínculos significativos entre a quantidade de mensagens positivas e negativas publicadas nas redes sociais e o desempenho financeiro dos filmes que eram citados nessas mensagens. Krauss, Nann e Simon (2008) já demonstraram ser possível prever indicados ao *Oscar* e encontrar correlações entre o desempenho financeiro de filmes e suas respectivas avaliações em fóruns *online*, como o *Internet Movie Database – IMDb*[3].

Este estudo visa aplicar técnicas da área de Análise de Sentimentos em uma base de dados de *tweets* referentes aos filmes indicados a Melhor Filme no *Oscar 2017*. Após aplicar técnicas de Processamento de Textos e Análise de Sentimentos, será possível descobrir a opinião dos usuários sobre os filmes e se essas opiniões têm alguma relação com o resultado da premiação.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo deste trabalho é verificar se existe uma correlação entre os vencedores do *Oscar 2017* e o sentimento expresso pelas pessoas por meio da rede social *Twitter* em relação aos filmes indicados. Para isso, serão usadas técnicas de pré-processamento e Análise de Sentimentos em uma base de dados composta por *tweets*, escritos na língua inglesa, referentes aos filmes indicados à categoria de Melhor Filme do *Oscar 2017*.

Serão utilizadas e comparadas diferentes técnicas de pré-processamento de texto, bem como diferentes algoritmos de classificação de texto, que indicam se o sentimento expresso no *tweet* em relação a um filme é considerado “positivo”, “negativo” ou “neutro”.

[3] http://www.imdb.com

Ao fim, será analisada a correlação entre os resultados dessa classificação e o resultado do *Oscar 2017*.

1.2.2 Objetivo específico

- Construir uma base rotulada para ser utilizada em experimentos, classificando, manualmente, uma certa quantidade de *tweets* – escolhidos de forma aleatória – com o sentimento "positivo", "negativo" ou "neutro" em relação a um dado filme. A base será construída a partir da coleta de *tweets* referentes aos indicados à categoria de Melhor Filme do *Oscar 2017* e publicados no período compreendido entre a data de divulgação dos indicados ao *Oscar 2017* (24/01/2017) e um dia antes da premiação (25/02/2017).
- Analisar algoritmos de classificação da literatura que usem diferentes abordagens para classificar os *tweets*, escolhendo o que obtém melhor desempenho para o problema.
- Sumarizar os sentimentos relativos aos indicados ao *Oscar 2017* na categoria de Melhor Filme e comparar essa sumarização com o resultado da premiação.
- Investigar se a opinião da Academia de Artes e Ciências Cinematográficas sobre os filmes em questão tem alguma relação com a opinião do público do *Twitter*, com a intenção de descobrir se é possível predizer quais serão os ganhadores ou perdedores da premiação.

1.3 Organização do Trabalho

O restante deste trabalho é organizado da seguinte maneira.

- **Capítulo 2 — Revisão Bibliográfica:** a fundamentação teórica necessária para o desenvolvimento e entendimento do estudo é apresentada, incluindo uma visão geral sobre redes sociais e sua importância no mundo contemporâneo, além da descrição de conceitos de Análise de Sentimentos e citação de trabalhos relacionados.
- **Capítulo 3 — Desenvolvimento:** expõe os métodos utilizados para desenvolvimento do trabalho, descrevendo cada uma das etapas a se executar com o objetivo de observar se há alguma relação entre o sentimento expresso nos *tweets* publicados sobre os filmes indicados à categoria de Melhor Filme do *Oscar 2017* e o resultado da premiação.
- **Capítulo 4 — Resultados:** os dados obtidos a partir da classificação realizada com o classificador escolhido são observados e analisados em comparação com o

resultado do *Oscar 2017*. Vários *rankings* baseados nesses resultados são construídos e comparados com o *ranking* da premiação.

- **Capítulo 5 — Conclusão:** as principais conclusões e contribuições deste trabalho são apresentadas, além de sugestões de trabalhos futuros.

2 Revisão Bibliográfica

O objetivo deste capítulo é apresentar a fundamentação teórica necessária para o desenvolvimento e entendimento deste estudo, além de trabalhos relacionados ao tema.

A Seção 2.1 apresenta uma visão geral sobre redes sociais e a sua importância no mundo atual, principalmente no âmbito do *marketing*, de manifestações sociais e da indústria de entretenimento. O *Twitter* é descrito mais detalhadamente por ser a rede social foco deste trabalho.

A Seção 2.2 descreve a Análise de Sentimentos e os desafios que envolvem a realização de uma tarefa nessa área utilizando o *Twitter* como fonte de dados.

A Seção 2.3 expõe as etapas referentes à tarefa de Análise de Sentimentos e como podem ser executadas. São elas: coleta de dados, pré-processamento dos dados, construção da base de dados rotulada, classificação dos textos e validação dos resultados da classificação. Formas de analisar a correlação entre o sentimento dos *tweets* e o objetivo alvo também são descritas.

A Seção 2.4 cita trabalhos relacionados e suas principais colaborações para área de Análise de Sentimentos.

2.1 Redes sociais

As redes sociais são espaços na *Internet* onde diferentes entidades (usuários, grupos ou organizações) são capazes de criar e compartilhar diversos tipos de conteúdo, além de acessar as publicações das outras entidades na rede (TEIXEIRA; AZEVEDO, 2011). Essas redes sociais têm atraído milhões, ou até bilhões de usuários, que representam mais de dois terços da população *online* global (BENEVENUTO; ALMEIDA; SILVA, 2011). Algumas redes sociais são citadas a seguir.

O *Facebook*[1] foi criado em 2004 como uma rede social reservada para os estudantes da Universidade de Harvard e hoje atua mundialmente, contando com uma média diária de 1,28 bilhão de usuários ativos e tem mais de quarenta escritórios espalhados pelo mundo (FACEBOOK, 2017). Alguns de seus objetivos são possibilitar às pessoas o contato com amigos e familiares, a expressão de suas opiniões e a descoberta de informações sobre o mundo.

Já o *Instagram*[2] e o *Snapchat*[3] são exemplos de aplicativos focados no compartilha-

[1] http://www.facebook.com
[2] http://www.instagram.com
[3] http://www.snapchat.com

mento de fotos e vídeos. Com menos de sete anos no ar, o *Instagram* possui cerca de 700 milhões de usuários ativos mensalmente (INSTAGRAM, 2017) enquanto que os adeptos ao *Snapchat* somaram aproximadamente 3 bilhões de *snaps*[4] enviados diariamente por meio da rede em Maio de 2017 (SMITH, 2017).

O *Twitter* é uma rede social que tem uma proposta incomum, limitando as postagens de seus usuários a textos com até 140 caracteres. Essa rede será descrita em detalhes na subseção 2.1.1, visto que foi escolhida como fonte de dados para este estudo.

2.1.1 O *Twitter*

A rede social *Twitter*[5] foi lançada em 2006 com a proposta excepcional de que cada postagem, chamada de *tweet*, não deveria exceder o limite de 140 caracteres[6]. Devido à sua crescente popularidade, a rede logo se tornou mais eficiente e acessível aos seus usuários, já que o seu principal objetivo é o intercâmbio de ideias e informações entre seus usuários, relatando acontecimentos mundiais em tempo real (TWITTER, 2017a).

Ao criar uma conta na rede, é gerado um perfil similar a um *microblog*, que apresenta os *tweets* publicados pela pessoa, que pode seguir outros perfis e também ser seguida. Ao acessar o seu *feed* de notícias, o usuário visualiza as postagens dos perfis que segue e, da mesma forma, suas atualizações são exibidas no *feed* de seus seguidores.

Alguns diferenciais do *Twitter* são *hashtags* – palavras-chave precedidas do caractere #, indicando que o *tweet* faz referência a um certo tópico – e *retweets* – indicação de que um usuário concorda com o *tweet* de outro, replicando a postagem em seu perfil (FILHO, 2014). Com mais de 313 milhões de usuários ativos mensalmente (TWITTER, 2017a), o *Twitter* continuamente sofre alterações para manter seus usuários ativos na rede, adaptando-se às suas necessidades. Atualmente, algumas das alterações mais significativas à interface da rede foram a inclusão de imagens aos *tweets*, criação de enquetes a serem respondidas por seguidores e incorporação do botão "coração" (TWITTER, 2017b), similar ao "curtir", do *Facebook* – ambos indicam que o seguidor gostou da publicação.

2.1.2 *Marketing* nas redes sociais

Apesar de serem usadas principalmente para lazer e entretenimento, empresas viram nas redes sociais a oportunidade de divulgarem seus produtos e serviços. De acordo com Patel (2015), por meio de redes sociais o *marketing* se torna mais dinâmico e pessoal, proporcionando à organização alcançar seu público alvo diretamente por meio de uma

[4] *Snap* é o termo utilizado para designar fotos e vídeos compartilhados através do *Snapchat*.
[5] http://www.twitter.com
[6] A partir de Novembro de 2017, o *Twitter* expandiu para 280 o limite de caracteres de um *tweet* (http://bit.ly/TwLim). Porém, a coleta de dados para este trabalho foi realizada antes do estabelecimento do novo limite na rede social, dessa forma, o limite original será considerado em outras seções deste trabalho.

rede específica, de acordo com a localização geográfica ou filtrando o público a partir dos perfis e grupos que os usuários acompanham.

As redes sociais têm sido bastante usadas para analisar a aceitação do público em relação a um produto, serviço ou acontecimento, uma vez que análise do sentimento de satisfação ou insatisfação expresso pelas frases que os clientes publicam nas rede sociais torna possível que empresas saibam sobre a reputação de seus produtos ou serviços de acordo com os usuários das redes *online* (FELIX, 2016). Teixeira e Azevedo (2011) mostram que essa abordagem também pode ser expandida para a área de entretenimento, posto que encontraram correlações significativas entre o número de mensagens positivas e negativas sobre um filme nas redes sociais e os respectivos valores de bilheteria para cada filme.

Dessa forma, as áreas de Mineração de Dados e Análise de Sentimentos se tornam essenciais para o *marketing* bem sucedido nas redes sociais. A cada minuto, cerca de 350 mil *tweets* são publicados (STATS, 2017) e neles existem informações relevantes que permitem às organizações descobrir a opinião de seus seguidores em relação aos seus produtos, serviços e à própria companhia em si (FILHO, 2014).

2.1.3 Manifestações nas redes sociais

Por estarem bastante presentes no cotidiano de seus usuários, algumas redes sociais como o *Twitter* e o *Facebook* se tornaram importantes na troca de informações sobre manifestações políticas e catástrofes naturais.

No Brasil, as redes sociais se tornaram fundamentais na divulgação de protestos contra o Estado. Em 2013, *Twitter* e *Facebook* foram usadas para divulgar protestos contra o aumento das tarifas de ônibus (COSTA, 2013), e em 2015 e 2016, para informar sobre as manifestações contra e a favor do *impeachment* da ex-presidente Dilma Rousseff (G1, 2016) (ROSSI, 2016).

Mandel et al. (2012) apresentaram o *Twitter* como uma fonte ao vivo de percepções públicas de um desastre natural. A partir da Análise de Sentimentos expressos em *tweets* relacionados ao furacão Irene, que atingiu o Caribe e a costa leste dos Estados Unidos em Agosto de 2011, os autores foram capazes de apurar diferentes aspectos demográficos da população afetada.

Vítimas de enchentes também já foram beneficiadas graças às redes sociais, tornando possível a realização de assistência remota por equipes de resgate [7]. Além disso, as diferentes redes sociais foram de suma importância para refugiados da Síria, que puderam enviar informações sobre suas condições em tempo real [8].

[7] http://www.ndtv.com/tamil-nadu-news/chennai-floods-ndrf-uses-social-media-to-reach-out-to-people-1251104

[8] http://bit.ly/SyrianBfeed

2.1.4 Indústria do entretenimento nas redes sociais

A indústria cinematográfica é uma das que se beneficia das redes sociais para realizar *marketing* de seus produtos e acompanhar o perfil e gostos de seus clientes. Filmes como *A Bela e a Fera* e *Homem-Aranha: De Volta ao Lar*, ambos lançados em 2017, tiveram forte divulgação *online*, por meio de suas respectivas páginas em redes sociais como *Twitter* e *Facebook*. O primeiro filme atraiu mais de 17 milhões de seguidores no *Facebook*[9], enquanto o segundo ultrapassou a marca de 20 milhões[10]. No Twitter, os números também impressionam, com *Homem-Aranha* acumulando mais de 560 mil fãs [11] e *A Bela e a Fera* reunindo mais de 177 mil seguidores [12]. Além da imensa visibilidade ocasionada pelas redes sociais, nelas as distribuidoras de filmes também têm acesso a uma ampla quantidade de opiniões que podem influenciar nos próximos lançamentos e na forma como serão divulgados.

Artistas musicais e suas gravadoras também desfrutam do poder das redes sociais para divulgarem suas músicas e descobrirem novas formas de atraírem fãs. Em Junho de 2017, a cantora Katy Perry transmitiu ao vivo os acontecimentos de sua vida durante quatro dias usando a rede social de compartilhamento de vídeos *YouTube*, com o objetivo de promover o lançamento de seu álbum. Essa transmissão acabou atraindo mais de 49 milhões de pessoas de 190 países, que também eram capazes de realizar comentários em tempo real por meio do *YouTube* [13].

As companhias do mercado editorial também se aproveitam das redes sociais para fins de divulgação e para receberem retorno de seus leitores. A Editora Intrínseca[14] e o Grupo Editorial Record[15] são exemplos que utilizam do *Facebook*, *Twitter* e *Snapchat* para se aproximarem de seus clientes.

2.2 Análise de Sentimentos

A Análise de Sentimentos, também chamada de Mineração de Opiniões, é definida por Liu (2012) como a área de estudo que analisa as opiniões, sentimentos, avaliações e atitudes das pessoas com relação a diferentes entidades – que podem ser produtos, serviços, organizações, individuais, eventos, tópicos, problemas e seus respectivos atributos. O significado do termo "opinião" ainda é muito amplo, todavia a Análise de Sentimentos evidencia principalmente opiniões que expressam ou implicam sentimentos positivos ou

[9] https://www.facebook.com/DisneyBeautyAndTheBeast
[10] https://www.facebook.com/HomemAranha
[11] https://twitter.com/SpiderManMovie
[12] https://twitter.com/beourguest
[13] http://www.billboard.com/articles/columns/pop/7833673/katy-perry-witness-world-wide-49-million-views
[14] http://www.intrinseca.com.br
[15] http://www.record.com.br

negativos.

As redes sociais *online* se tornaram espaços essenciais para que os usuários da *Internet* possam consumir informações e expressar suas opiniões por meio de textos, fotos, vídeos e músicas. Essa disseminação de ideias acabou chamando a atenção de estudiosos e de empresas que procuram encontrar dados relevantes dentre a imensa quantidade de informação que é gerada a cada momento na rede (BOTHOS; APOSTOLOU; MENTZAS, 2010). Dessa forma, é comum que usuários busquem por opiniões sobre um produto ou serviço em *reviews*, *blogs* e redes sociais para serem auxiliados na decisão de realizar a compra ou confiar no serviço oferecido por uma empresa (OLIVEIRA, 2013). Consequentemente, as áreas de Mineração de Dados e Análise de Sentimentos têm recebido grande atenção não só de cientistas de dados, mas também de empresas que têm forte presença *online*.

Ainda que a Análise de Sentimentos nas redes sociais seja uma área bastante interessante e que tem despertado o interesse de empresas e pesquisadores, para a realização dessa tarefa, diversos obstáculos precisam ser superados, como por exemplo, a detecção de diferentes entidades e seus relacionamentos, a grande quantidade de lixo e ruído informacional presentes nas postagens (BENEVENUTO; ALMEIDA; SILVA, 2011) e, claro, a grande quantidade de dados (FELIX, 2016).

2.2.1 Desafios em Análise de Sentimentos no *Twitter*

O *Twitter* é uma das redes que encoraja a publicação de textos curtos e, por isso, dificuldades podem surgir durante o processamento dos *tweets*. Algumas delas são relevantes a este estudo e são citadas a seguir.

- Ambiguidade: algumas palavras têm múltiplos sentidos, podendo inclusive expressar sentimentos opostos (LIU, 2012), como pode ser exemplificado a partir do uso da palavra "bad" nos *tweets* sobre o filme *La La Land*, mostrados nas Figuras 1 e 2.

Figura 1 – *EU QUERO MUITO ASSISTIR LA LA LAND*: "bad" ("muito") expressa sentimento positivo.

Figura 2 – *la la land é ruim*: "bad" ("ruim") expressa sentimento negativo.

- Títulos de filmes, produtos, serviços, etc. que são palavras do cotidiano: *Arrival*, *Fences*, *La La Land*, *Lion* e *Moonlight* são exemplos de filmes com títulos que remetem a palavras utilizadas no cotidiano das pessoas que falam a língua inglesa. Em uma aplicação envolvendo a Análise de Sentimentos relacionada a filmes, uma busca de *tweets* contendo alguma dessas palavras-chave pode retornar resultados que não tenham qualquer relação com o filme em questão, como é exemplificado na Figura 3.

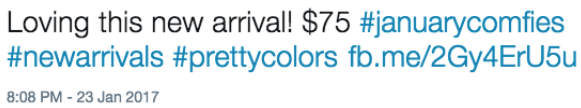

Figura 3 – Exemplo de *tweet* coletado em uma consulta pelo filme *Arrival*, mas que indica a chegada de um novo produto a uma loja de roupas.

- Sarcasmo: sentenças sarcásticas como "What an amazing movie! I just lost two hours that I'll never get back" dificultam a detecção do sentimento, pois o sentido acaba sendo o oposto do que se espera (OLIVEIRA, 2013).

- Variações na ortografia: por serem mensagens curtas em um contexto informal, muitas vezes faz-se necessário que os autores de *tweets* utilizem de gírias e abreviações para se expressarem em suas postagens, como por exemplo o *tweet* mostrado na Figura 4.

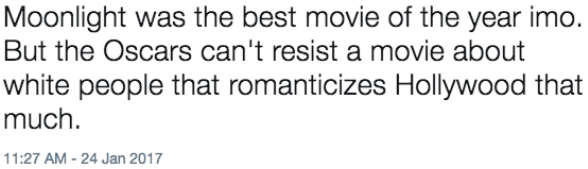

Figura 4 – *Tweet* em que o autor usa "imo" para abreviar "in my opinion".

- Outra prática comum, que ocorre em cerca de um a cada seis *tweets* (FELIX, 2016), é o uso da repetição de letras para indicar intensidade ou enfatizar um sentimento, como mostrado na Figura 5.

Figura 5 – *Tweet* que contém repetição de letras para indicar intensidade do sentimento.

2.3 Etapas da Análise de Sentimentos

Por ser uma tarefa complexa, a Análise de Sentimentos é dividida em etapas, que serão descritas a seguir.

2.3.1 Coleta de dados

Primeiramente, é importante definir a fonte de dados a ser utilizada, ou seja, a rede social escolhida, e a forma como os dados serão coletados. O *Facebook* oferece uma API[16] para auxiliar nesta tarefa, assim como o *Twitter*[17], apesar de a última ser bastante limitada, por exemplo, restringindo a quantidade de resultados e retornando apenas *tweets* publicados há no máximo sete dias de acordo com a data que a consulta é realizada.

Muitos pesquisadores preferem criar seus próprios robôs para coleta de dados ou utilizar outros disponíveis *online*, como o *GetOldTweets*[18], que supera algumas limitações da API oficial do *Twitter*, como por exemplo, não restringir o intervalo de tempo da busca por *tweets*. Ao realizar a recuperação de *tweets* a partir de certa palavra-chave, para cada instância retornada, além do texto do *tweet*, também são retornados alguns metadados do mesmo, como data, autor, número de *retweets*, número de *likes*, entre outros.

2.3.2 Construção da base de dados rotulada

Os dados obtidos a partir da coleta de dados compõem a base de dados bruta a ser utilizada no experimento. Em Análise de Sentimentos, a tarefa de classificar cada um dos *tweets* em relação ao sentimento expresso neles pode envolver o uso de técnicas de Aprendizagem de Máquina. Uma delas é o Aprendizado Supervisionado, que consiste no uso de um conjunto de treinamento criado para treinar o classificador (BAEZA-YATES; RIBEIRO-NETO, 2013).

Para tal, é importante construir uma base de dados rotulada a partir de uma amostra da base original, a fim de usá-la como conjunto de treinamento. A tarefa de rotular a base é trabalhosa, exige bastante tempo e cuidado, além de ser uma atividade que é muitas vezes desempenhada por mais de um especialista humano (De Smedt; DAELEMANS, 2012).

Não existe um padrão com relação ao tamanho da base rotulada a ser utilizada em experimentos focados nas redes sociais, uma vez que cada aplicação exige uma amostra diferente (ARAÚJO et al., 2013) (REIS; GONÇALVES; ARAÚJO, 2012). Porém, é importante garantir que haja uma quantidade notável de instâncias rotuladas para cada uma

[16] https://developers.facebook.com/docs/graph-api
[17] https://dev.twitter.com/rest/public
[18] https://github.com/Jefferson-Henrique/GetOldTweets-python

das classes em questão e assegurar que a base como um todo seja representada (ARAÚJO et al., 2013).

2.3.3 Pré-processamento dos dados

A etapa de pré-processamento dos dados é de fundamental importância para a Análise de Sentimentos. Ela tem como objetivo descartar o que é considerado irrelevante à etapa de classificação e/ ou alterar o formato dos dados de forma a auxiliar nessa etapa. O pré-processamento é composto por diversas tarefas, que variam em conformidade com as particularidades dos dados que estão sendo processados, isto é, cada fonte de dados (exemplo: redes sociais) exige o uso de diferentes tarefas para que um resultado satisfatório seja obtido.

Algumas tarefas relacionadas ao pré-processamento de dados obtidos a partir do *Twitter* são apresentadas a seguir, bem como exemplos de trabalhos que as adotaram.

- Remoção de *links*, pois esses termos não possuem conteúdo semântico (FILHO, 2014) (FELIX, 2016) (SOUZA, 2012) (ALMEIDA, 2012) (GO; BHAYANI; HUANG, 2009).
- Remoção de caracteres não alfabéticos e pontuação, já que não agregam valor à classificação (FILHO, 2014) (FELIX, 2016).
- Substituição de caracteres acentuados pelos correspondentes sem acentuação, com o objetivo de padronizar o texto (FILHO, 2014) (SOUZA, 2012)
- Remoção de repetição de letras: uma prática comum entre usuários de redes sociais é repetir letras para dar maior ênfase em seu sentimento, por exemplo, se uma pessoa gostou muito de um filme, ela pode escrever *"this movie is really gooooood!"* (FILHO, 2014) (RIBEIRO, 2015) (GO; BHAYANI; HUANG, 2009).
- Remoção de citação a outros usuários: no *Twitter*, usa-se o símbolo "@" para citar outros usuários da rede social, por exemplo *@john olá! tudo bem?* (FILHO, 2014) (FELIX, 2016) (ALMEIDA, 2012) (GO; BHAYANI; HUANG, 2009).
- Conversão de letras maiúsculas em minúsculas, a fim de padronizar o texto (FELIX, 2016).
- Substituição de *emoticons* por palavras correspondentes, com a intenção de enriquecer o modelo de classificação (FELIX, 2016).
- Substituição de gírias e abreviações por expressões completas. Assim, um maior número de palavras e expressões pode ser reconhecido pelo classificador (OLIVEIRA, 2013) (RIBEIRO, 2015) (SOUZA, 2012).

- Remoção de *stop words*, palavras que são bastante comuns em um idioma e, portanto, não possuem muito valor semântico. Alguns exemplos de *stop words* da língua inglesa são *"the"*, *"with"* e *"you"* (FELIX, 2016) (RIBEIRO, 2015) (FILHO, 2014) (SOUZA, 2012) (SCHMITT, 2013).
- *Stemming*, que consiste na separação de cada termo em radical e terminação, removendo-se a terminação. Dessa forma, palavras como *"computer"*, *"computing"* e *"computed"* seriam todas reduzidas a *"comput"* após *stemming*. Isso resultaria em um maior peso para esse termo, mas também poderia anular o sentido original da palavra (FELIX, 2016) (OLIVEIRA, 2013) (RIBEIRO, 2015).
- Remoção de termo da busca, já que é constante em todos os resultados (ALMEIDA, 2012).

Essas tarefas podem ser realizadas a partir da criação de algoritmos próprios ou utilizando *softwares* adequados para realização de tarefas de Mineração de Dados e Aprendizagem de Máquina, como *NLTK* (LOPER; BIRD, 2002) e *Weka* (SMITH; FRANK, 2016).

Ao fim da etapa de pré-processamento, tem-se uma base de dados mais consistente que pode ser então usada no processo de classificação.

2.3.4 Classificação dos textos

É a etapa em que algoritmos de classificação da literatura são utilizados para classificar um texto, ou seja, atribuir um sentimento como "positivo", "neutro" ou "negativo". Diferentes trabalhos da literatura propõem variados métodos de classificação, porém poucos são os estudos que tentam analisar a influência desses métodos na tarefa de Análise de Sentimentos (RIBEIRO et al., 2016). Abordagens como aprendizado supervisionadas e não-supervisionado têm sido usadas na tarefa de classificação (FELIX, 2016).

Segundo Baeza-Yates e Ribeiro-Neto (2013), um algoritmo de classificação é considerado supervisionado quando um conjunto de treinamento é usado para treinar o classificador. Para os algoritmos não supervisionados, não é necessário que sejam fornecidos exemplos de treinamento (BAEZA-YATES; RIBEIRO-NETO, 2013).

A seguir, são descritas as estratégias de classificação que serão foco deste trabalho.

2.3.4.1 *Aprendizado supervisionado*

De acordo com Baeza-Yates e Ribeiro-Neto (2013), um algoritmo de classificação é dito supervisionado caso um conjunto de treinamento criado por humanos seja usado para treinar o classificador – esse conjunto pode ser representado por uma base de dados rotulada, como explicado na subseção 2.3.2.

O algoritmo *Naive Bayes* é um dos métodos de aprendizado supervisionado mais utilizados no escopo de análise de sentimentos devido ao seu desempenho notável na classificação de textos e análise de sentimentos (RIBEIRO, 2015). É um algoritmo probabilístico que se baseia no conhecimento prévio do problema, combinado a exemplos de treinamento, para determinar a probabilidade de um documento pertencer a certa classe (BAEZA-YATES; RIBEIRO-NETO, 2013) (SCHMITT, 2013).

O classificador *Naive Bayes* é fundamentado no Teorema de Bayes, representado pela Equação 2.1, definida como a "probabilidade de A dado B", isto é, determina a probabilidade da hipótese A estar em B, dado um conjunto de evidências B (SCHMITT, 2013).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

Esse classificador é considerado ingênuo (*naive*) por assumir que os termos de uma instância são condicionalmente independentes entre si, não exercendo influência um sobre o outro. Os termos apenas influenciam a classe a que se remetem (SCHMITT, 2013).

É importante enfatizar que algoritmos de aprendizado supervisionado como o *Naive Bayes* necessitam de uma base de dados rotulada para ser usada durante a fase de treinamento (BAEZA-YATES; RIBEIRO-NETO, 2013).

Os classificadores Bayesianos podem ser abordados de diferentes formas, algumas delas sendo os modelos binário e multinomial (BAEZA-YATES; RIBEIRO-NETO, 2013). No modelo binário, cada documento é representado por um vetor binário e a presença ou ausência de um termo são representadas pelos valores 1 ou 0, respectivamente (SCHMITT, 2013). No modelo multinomial, cada documento é representado por um vetor de inteiros, indicando o número de vezes que cada termo acontece no documento (SCHMITT, 2013).

2.3.4.2 *Aprendizado por supervisão à distância*

O aprendizado por supervisão à distância utiliza de uma forma alternativa para geração de dados de treinamento. Nessa estratégia, utiliza-se uma base de dados já existente para coletar instâncias com vínculo à relação que deseja-se analisar. Em seguida, essas instâncias são utilizadas para gerar automaticamente conjuntos de treinamento (DEEPDIVE, 2017).

Sentiment140 é uma ferramenta para análise de sentimentos específica para o *Twitter* que utiliza aprendizado por supervisão à distância. Ao acessar seu *website*[19], é possível entrar com uma palavra-chave e descobrir o sentimento relacionado a ela de

[19] http://www.sentiment140.com/

acordo com os usuários do *Twitter*, dessa forma, sendo possível analisar o que os usuários pensam sobre uma marca, produto ou tópico.

Uma API[20] do *Sentiment140* é disponibilizada para propósitos acadêmicos e permite que vários *tweets* sejam classificados em massa. A polaridade é retornada quando uma sentença é fornecida como entrada. O valor "0" indica que aquela sentença é negativa, "2" designa neutralidade e "4" classifica uma sentença como positiva.

Esta ferramenta utiliza de aprendizado por supervisão à distância e um classificador de Máxima Entropia[21] para calcular a polaridade de uma sentença a partir de uma base de dados rotulada de acordo com os *emoticons* presentes nos *tweets* (GO; BHAYANI; HUANG, 2009) que a compõem. Essa base é utilizada como conjunto de treinamento (GO; BHAYANI; HUANG, 2009).

O classificador de Máxima Entropia favorece os modelos mais uniformes que satisfaçam certa regra. Em um cenário de duas classes, utiliza de regressão logística para encontrar uma distribuição sobre as classes e, diferente do *Naive Bayes*, não assume que as instâncias sejam independentes entre si (GO; BHAYANI; HUANG, 2009).

De acordo com Go, Bhayani e Huang (2009), o método do *Sentiment140*, que utiliza *tweets* com *emoticons* como base de treino, provou-se como uma boa maneira para classificar *tweets*, uma vez que algoritmos de classificação como *Naive Bayes*, Máxima Entropia e Máquina de Vetor de Suporte (*Support Vector Machine*) obtiveram ótimos índices de acurácia quando testados.

2.3.4.3 *Função de polaridade*

TextBlob[22] é uma biblioteca para processamento de texto escrita para a linguagem *Python* que fornece uma API para solução de diversas tarefas relacionadas ao processamento de linguagem natural (LORIA et al., 2017). Essa ferramenta realiza integração com a plataforma NLTK (*Natural Language Toolkit*)[23] e com o módulo de mineração *Web Pattern*[24].

Uma das formas de se classificar texto com o *TextBlob* é utilizando a função *polarity*, que retorna a polaridade de uma sentença dada como entrada. Essa função utiliza a mesma implementação fornecida pelo módulo *Pattern*, classificando as sentenças de acordo com um léxico próprio construído por especialistas e manualmente rotulado de acordo com a força de polaridade, subjetividade e intensidade de cada palavra (De Smedt; DAELEMANS, 2012). Esse léxico é um dicionário composto de adjetivos frequentes em avaliações de produtos *online* (De Smedt; DAELEMANS, 2012).

[20] http://help.sentiment140.com/api
[21] Ver nota de rodapé 20
[22] http://textblob.readthedocs.io/en/dev/
[23] http://www.nltk.org/
[24] https://www.clips.uantwerpen.be/pattern

Métodos de aprendizagem de máquina foram utilizados para expandir o dicionário com novos adjetivos. Para avaliar o método, textos de avaliações de produtos foram classificados e comparados com a classificação de estrelas atribuída a cada um por consumidores. Segundo De Smedt e Daelemans (2012), a obtenção da polaridade de sentenças a partir desse léxico devolveu resultados promissores.

O valor de polaridade obtido a partir da função *polarity* é um número entre "-1" e "1" (LORIA et al., 2017), que é resultado de um cálculo da intensidade média das palavras da sentença que também estão presentes no léxico (De Smedt; DAELEMANS, 2012). Valores de polaridade mais próximos de "1" sugerem maior intensidade de sentimento positivo para a sentença, enquanto que valores mais próximos de "-1", apontam para maior intensidade de sentimento negativo na sentença; os valores próximos de "0" indicam neutralidade. (LORIA et al., 2017).

O léxico proposto por De Smedt e Daelemans (2012) é constituído de palavras em Holandês, mas uma versão em Inglês também foi disponibilizada posteriormente e é a utilizada pela função do *TextBlob*.

2.3.5 Validação dos resultados da classificação

Uma vez construído o modelo de classificação, é importante analisar sua capacidade preditiva usando diferentes amostragens dos dados e diferentes medidas de validação.

Uma das estratégias de amostragem dos dados conhecidas é a validação cruzada de 10 pastas (*10-fold cross-validation*), que consiste na divisão da base de dados em 10 partes aleatórias, que são testadas individualmente, utilizando as outras nove partes restantes como conjunto de treinamento (GIGLIOTTI, 2012).

Após os testes realizados com os classificadores utilizando a base de dados rotulada como conjunto de treinamento, é possível obter as seguintes medidas, considerando um cenário em que o classificador é treinado usando as classes "positivo", "negativo" e "neutro".

- **VP** (verdadeiros positivos): instâncias positivas corretamente classificadas.
- **VN** (verdadeiros negativos): instâncias negativas corretamente classificadas.
- **VE** (verdadeiros neutros): instâncias neutras corretamente classificadas.
- **FP** (falsos positivos): instâncias negativas ou neutras classificadas como positivas.
- **FN** (falsos negativos): instâncias positivas ou neutras classificadas como negativas.
- **FE** (falsos neutros): instâncias positivas ou negativas classificadas como neutras.

Com essas medidas, é possível criar uma matriz de confusão, que é uma forma de demonstrar e avaliar os resultados obtidos a partir de um algoritmo de classificação (SCHMITT, 2013). As linhas da matriz representam as instâncias reais das classes, enquanto as colunas indicam como as instâncias foram classificadas pelo algoritmo, como mostrado na Tabela 1.

classificado como				
pos	neg	neu		
VP	FN	FE	pos	
FP	VN	FE	neg	real
FP	FN	VE	neu	

Tabela 1 – Matriz de confusão para três classes: positivo, negativo e neutro

Após construir a matriz de confusão para o classificador, é possível avaliar o desempenho do mesmo por meio de medidas de qualidade, como a acurácia (Equação 2.2), que permite calcular a proporção de instâncias que foram classificadas corretamente (SCHMITT, 2013).

$$acurácia = \frac{VP + VN + VE}{VP + VN + VE + FP + FN + FE} \tag{2.2}$$

Uma vez que o cálculo da acurácia é realizado, é possível validar o classificador em questão e comparar seu desempenho com o de outros.

2.3.6 Análise da correlação entre o sentimento dos *tweets* e o objetivo alvo

Após a classificação do sentimentos dos *tweets* em "positivo", "negativo" ou "neutro", a próxima etapa consiste em estabelecer correlações entre esse sentimento e o objetivo alvo de estudo. A fim de tentar estabelecer essa correlação, diferentes estratégias são usadas, por exemplo, as descritas a seguir.

- Nuvem de palavras: representa graficamente a frequência de ocorrência dos termos contidos em certa base de dados a partir de uma nuvem composta por esses termos (FILHO, 2014). Quanto maior for a ocorrência do termo, maior será o seu tamanho na nuvem[25]. A Figura 6 mostra a nuvem de palavras relativa aos *tweets* recuperados para o filme *Manchester by the Sea*.

  Pode-se perceber que as palavras que mais apareceram dentre os *tweets* foram "oscar", "movie", "best" – em referência às seis categorias do *Oscar 2017* em que o filme foi indicado – e "affleck" – menção ao ator Casey Affleck, que estrelou o longa-metragem e que foi indicado à categoria de Melhor Ator.

[25] https://worditout.com/word-cloud/create

Figura 6 – Nuvem de palavras contidas nos *tweets* coletados para o filme *Manchester by the Sea* (após pré-processamento) — Gerada a partir do *WordItOut* (https://worditout.com/word-cloud/create)

- Mapas de calor: representam graficamente os dados de acordo com a intensidade em que acontecem em determinadas áreas (FILHO, 2014), por exemplo: áreas mais visualizadas por um usuário ao acessar certo site, áreas em que certa doença mais ocorre no mundo, entre outros.

  Em 2015, o *Twitter* gerou um mapa de calor animado[26] que mostra quais áreas geográficas do mundo tiveram mais pessoas publicando *tweets* sobre o *Oscar 2015* de acordo com os acontecimentos da premiação. A Figura 7 ilustra um desses momentos.

- Gráficos: auxiliam na visualização dos dados, podendo ser de vários tipos. Por exemplo: barras, pizza, radar, entre outros. A Figura 8 mostra um gráfico de linhas utilizado por Mandel et al. (2012) em seu estudo sobre o sentimento dos usuários do *Twitter* em relação ao furacão Irene.

2.4 Trabalhos relacionados

A quantidade de trabalhos relacionados a Análise de Sentimentos é abundante, por ser uma área que tem um objetivo excepcional, que é dar sentido à imensidão do *big data*. Cada vez mais esta tem sido uma área que gera retornos notáveis, o que inspirou

[26] http://bit.ly/Oscar2015HeatMap

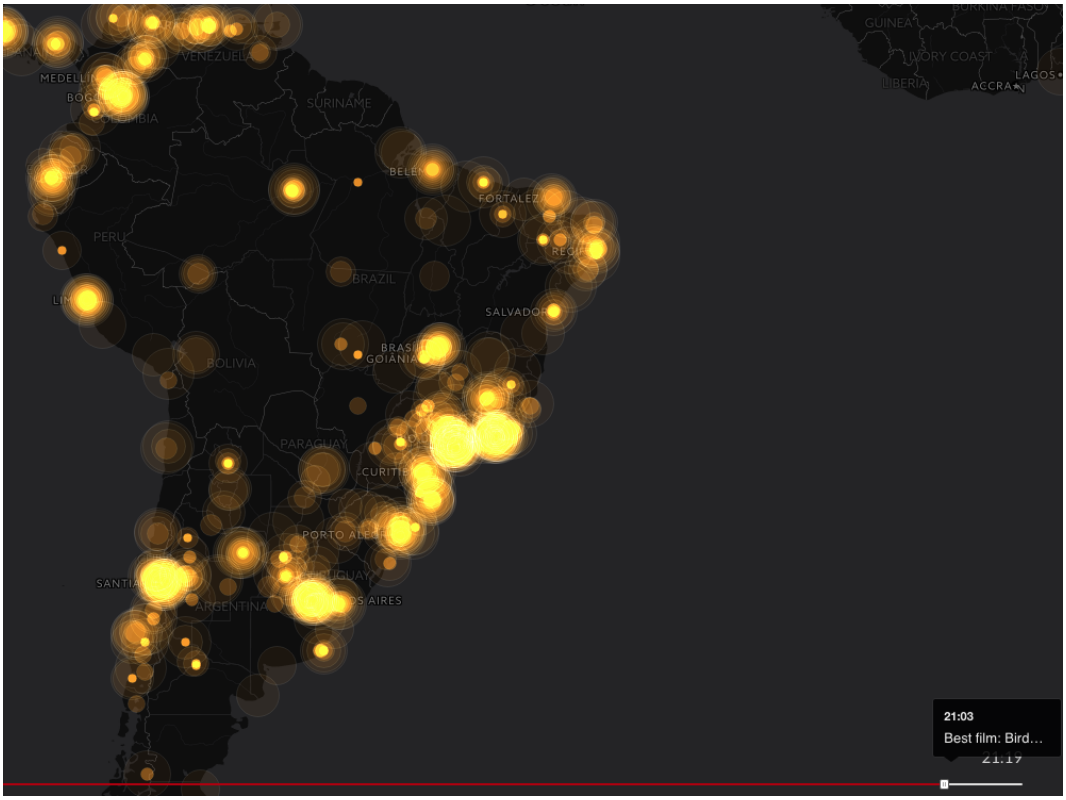

Figura 7 – Mapa de calor dos *tweets* publicados durante o anúncio do vencedor na categoria de Melhor Filme do *Oscar 2015* — Fonte: Twitter (http://bit.ly/Oscar2015HeatMap)

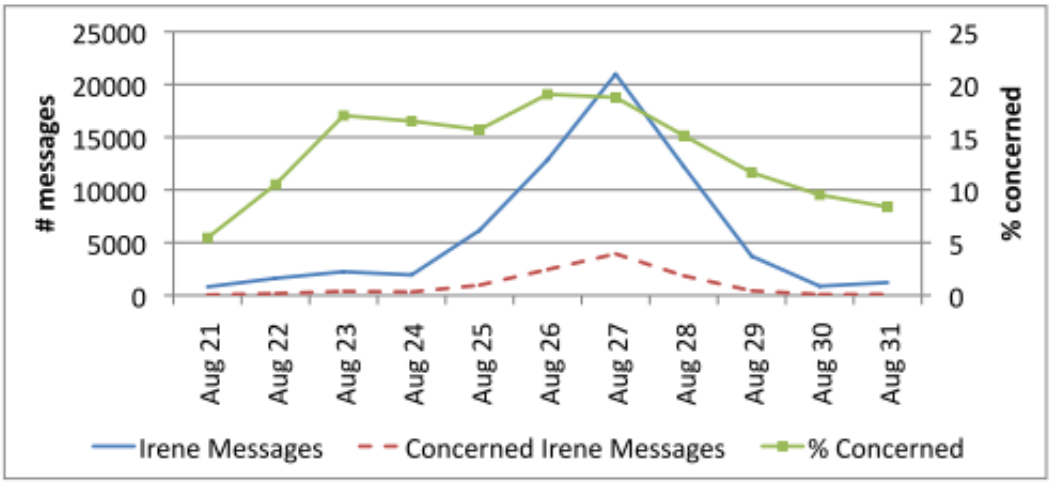

Figura 8 – Número de *tweets* publicados sobre o furacão Irene, além da quantidade e porcentagem de instâncias classificadas como "preocupado" ("*concerned*") pelo classificador utilizado (MANDEL et al., 2012)

pesquisadores como Benevenuto, Almeida e Silva (2011) e Cervi (2008) a explorarem os diferentes métodos e técnicas da área que podem ser aplicados às redes sociais.

O *Twitter* é uma das redes sociais mais utilizadas neste tipo de estudo por ser uma fonte abundante de opiniões pessoais vindas do mundo inteiro (PAK; PAROUBEK,

2010). Além disso, o número de postagens em texto publicadas na plataforma aumenta a cada momento e seu público é bastante heterogêneo, sendo formado por múltiplos grupos sociais com diferentes interesses (PAK; PAROUBEK, 2010).

Vários trabalhos têm sido desenvolvidos no intuito de correlacionar o sentimento positivo ou negativo dos *tweets* com relação a algum acontecimento. Em Filho (2014), o autor utilizou uma implementação do algoritmo de classificação *Naive Bayes* para classificar *tweets* relacionados aos jogos do Brasil na Copa do Mundo de 2014. Desse modo, ele foi capaz de mapear o sentimento dos usuários do *Twitter* com relação aos esses jogos, demonstrando seus resultados por meio de mapas de calor e nuvens de palavras.

Já um estudo publicado por Almeida (2012) expôs a grande incidência de *cyberbullying* realizada contra professores no *Twitter*. Durante uma semana, o autor coletou *tweets* com referências a professores e, a partir da aplicação de um filtro de classificação Bayesiano na base de *tweets*, concluiu-se que a violência virtual contra professores é real e diária, posto que mais da metade dos *tweets* direcionados aos professores foi classificada com sentimentos negativos.

Considerando o cenário de entretenimento, em especial a Análise de Sentimentos em relação a filmes, poucos trabalhos foram encontrados na literatura. Teixeira e Azevedo (2011) extraíram do *Twitter* e do *Facebook* publicações relacionadas a filmes que ainda não haviam sido lançados nos cinemas. Os autores tinham como objetivo descobrir se as informações contidas nesses textos seria útil para estimar se os filmes, ou outros produtos, a serem lançados teriam sucesso financeiro. Depois de tratar as mensagens, os autores tiveram auxílio do recurso léxico *SentiWordNet*[27] para classificar as mensagens e então foi utilizada a correlação de *Spearman* para determinar a existência de uma relação entre os resultados obtidos e o valor de bilheteria para cada filme. Após análise dos resultados, encontraram vínculos significativos entre a quantidade de mensagens positivas e negativas publicadas nas redes sociais e o desempenho financeiro dos filmes que eram citados nessas mensagens.

Krauss, Nann e Simon (2008) extraíram comentários dos usuários do *Internet Movie Database – IMDb*[28] para analisar suas respectivas opiniões com relação a diversos filmes. Após a coleta dos dados, os autores criaram um dicionário customizado para medir os níveis de positividade em um texto e, dessa forma, demonstraram ser possível prever indicados ao *Oscar* e encontrar correlações entre o desempenho financeiro de filmes e suas respectivas avaliações em fóruns *online*, como o *IMDb*.

Cetinsoy (2017) utilizou a plataforma de aprendizagem de máquina *BigML* para tentar prever os vencedores de oito categorias do *Oscar 2017*. Para isso, o autor e sua equipe construíram um conjunto de dados bastante rico, com informações sobre filmes

[27] http://sentiwordnet.isti.cnr.it/
[28] http://www.imdb.com

indicados e vencedores do *Oscar* e de outras premiações importantes da indústria cinematográfica entre os anos de 2000 e 2016. Além disso, foram coletadas avaliações e críticas de usuários do *IMDb*.

Como resultado, os modelos desenvolvidos pelo autor conseguiram acertar os vencedores de cinco das oito categorias em questão, reforçando como um bom conjunto de dados e um modelo de aprendizagem de máquina adequado podem retornar resultados significativos.

2.5 Considerações finais

O *Twitter* e outras redes tornaram-se espaços onde pessoas comentam sobre filmes, músicas, livros e outros assuntos da indústria de entretenimento. Muitos usuários do *Twitter* também o utilizam para comentar sobre premiações, como *Oscar* e *Grammy*, eventos esportivos e tópicos de interesse mundial. Dessa forma, a Análise de Sentimentos torna-se uma ferramenta imprescindível na descoberta das opiniões de um grande número de pessoas em relação a esses tópicos, mesmo com alguns desafios, como ambiguidade e sarcasmo – que são complexos de se detectar em textos –, o uso de gírias e repetições de letras, dentre outros.

Por ser uma área que tem gerado resultados significativos e devido à riqueza de dados presentes nas redes sociais, há uma grande quantidade de trabalhos relacionados. Porém, percebeu-se que há uma certa ausência de trabalhos que analisem a relação entre redes sociais e filmes, em especial em relação ao *Oscar*. Alguns poucos abordam temas semelhantes, como Teixeira e Azevedo (2011), em que os autores foram capazes de encontrar vínculos significativos entre o sentimento dos usuários de redes sociais e o desempenho financeiro de filmes, enquanto Krauss, Nann e Simon (2008) conseguiram prever indicados ao *Oscar* a partir da análise de opiniões de usuários nas redes sociais.

Neste estudo serão analisados *tweets* sobre os filmes indicados à categoria de Melhor Filme do *Oscar 2017* e, a partir do uso de técnicas de Análise de Sentimentos, será possível descobrir os sentimentos dos usuários do *Twitter* em relação a cada filme. Para isso, a tarefa de Análise de Sentimentos será dividida em cinco etapas: coleta de dados, pré-processamento dos dados, construção da base de dados rotulada, classificação dos textos e, por fim, validação dos resultados.

3 Desenvolvimento

Este capítulo expõe os métodos utilizados para o desenvolvimento deste estudo, que tem como objetivo descobrir se há alguma relação entre os indicados e vencedores do *Oscar 2017* e o sentimento dos usuários do *Twitter* em relação aos filmes indicados à categoria de Melhor Filme.

A Seção 3.1 apresenta como foi realizada a coleta dos dados do *Twitter* para compor a base de dados utilizada para o estudo.

A Seção 3.3 lista as etapas de pré-processamento realizadas neste trabalho, que foram executadas com o objetivo de tratar os dados, descartando o que é irrelevante para a etapa de classificação.

A Seção 3.2 descreve como foi realizado o processo de rotular *tweets* com um dos sentimentos (“positivo”, “negativo” ou “neutro”), com o objetivo de construir a base rotulada para ser utilizada como conjunto de treinamento para algoritmos de aprendizado supervisionado.

A Seção 3.4 expõe as três abordagens de classificação da literatura utilizadas para classificar os *tweets* da base rotulada. São elas: aprendizado supervisionado, aprendizado por supervisão à distância e função de polaridade.

A Seção 3.5 discute a forma utilizada para medir o desempenho dos classificadores avaliados.

A Seção 3.6 explica a medida criada especialmente para este estudo com o objetivo de obter um *ranking* do *Oscar 2017*, baseado na quantidade de indicações e vitórias que cada filme analisado conquistou.

A Seção 3.7 enuncia as ferramentas utilizadas para comparar o resultado do *Oscar 2017* com os sentimentos dos *tweets*.

3.1 Coleta de dados

O *Twitter* foi utilizado para a realização da coleta de dados deste estudo. Essa rede social foi escolhida devido à sua grande popularidade (TWITTER, 2017a) e ao fato de que todos os *tweets* são curtos, contendo no máximo 140 caracteres, possibilitando que se saiba a opinião de várias pessoas sobre um certo assunto de forma rápida e sucinta.

A coleta de *tweets* foi feita considerando o período compreendido entre 24 de Janeiro de 2017 – data em que foram anunciados os indicados ao *Oscar 2017* – e 25 de Fevereiro de 2017 – um dia antes da cerimônia de entrega dos prêmios (DONNELLY,

2017).

Inicialmente, a coleta de *tweets* seria realizada utilizando a API[1] disponibilizada pelo *Twitter* em seu *website* oficial. Porém, após análise da documentação, foi constatado que a ferramenta não seria útil para este estudo, uma vez que apenas recupera *tweets* publicados em até sete dias antes da data de consulta e retorna apenas um número limitado de instâncias. Como a coleta de dados foi realizada um mês após a cerimônia do *Oscar*, foi necessário buscar por uma ferramenta que possibilitasse a recuperação de *tweets* mais antigos.

Assim, para este estudo foi adotada a *GetOldTweets*[2], uma ferramenta escrita na linguagem *Python* que supera as limitações de período impostas pela API do *Twitter*, permitindo que sejam recuperados *tweets* publicados na rede social em qualquer data. A partir de sua interface de linha de comando, é possível realizar consultas ao *Twitter* e recuperar *tweets* de acordo com os argumentos especificados, como mostrado na Figura 9.

```
python Exporter.py --querysearch "la la land lang:en" --since 2017-01-24 --until 2017-02-25
```

Figura 9 – Consulta ao *Twitter* realizada por meio da interface de linha de comando do *GetOldTweets*

Os argumentos especificados nas consultas realizadas para este trabalho são descritos a seguir.

- *querysearch*: palavras-chave a serem consideradas na busca por *tweets*, seguidas do parâmetro *lang:en*, que refina a consulta para *tweets* escritos apenas na língua inglesa.

  As palavras-chave utilizadas como parâmetro *querysearch* foram os títulos originais dos filmes analisados: “arrival”, “fences”, “hacksaw ridge”, “hell or high water”, “hidden figures”, “la la land”, “lion”, “manchester by the sea” e “moonlight”.

- *since*: indica a data inicial a ser considerada na consulta. A data “2017-01-24” foi utilizada como parâmetro.

- *until*: indica a data final a ser considerada na consulta. A data “2017-02-25” foi utilizada como parâmetro.

Escolheu-se buscar por *tweets* escritos em Inglês devido ao fato de que esse é o idioma mais usado no *Twitter*, de acordo com o estudo conduzido pela companhia Semiocast, que realiza pesquisas relacionadas a inteligência de dados e mídias sociais (SEMIOCAST,

[1] https://developer.twitter.com/en/docs
[2] https://github.com/Jefferson-Henrique/GetOldTweets-python

2011). Além disso, há maior disponibilidade de ferramentas para pré-processamento e Análise de Sentimento para a língua inglesa (FELIX, 2016). A partir dessa decisão, garantiu-se que a base de dados para este estudo fosse bastante rica, composta por um número significativo de *tweets*.

Para este trabalho, foram realizadas nove consultas. As palavras-chave utilizadas foram representadas pelos títulos originais (em Inglês) dos filmes indicados à categoria de Melhor Filme do *Oscar 2017*.

Após cada execução do programa, um arquivo era gerado, incluindo os *tweets* encontrados e seus respectivos metadados. Porém, para este estudo, apenas o texto de cada *tweet* foi considerado relevante. A quantidade de *retweets* foi desconsiderada para evitar que um peso extra fosse atribuído a *tweets* com muitos *retweets* (GO; BHAYANI; HUANG, 2009). A Tabela 8 apresenta a quantidade de *tweets* recuperados para cada filme.

3.2 Criação da base de dados rotulada

A classificação de textos utilizando o algoritmo *Naive Bayes* é avaliada neste estudo. Por ser um modelo supervisionado, tornou-se necessária a construção de uma base de *tweets* rotulados manualmente, a fim de obter um conjunto de treinamento para este trabalho. Assim, uma amostra aleatória de *tweets* foi selecionada a partir da base de dados coletada originalmente e cada *tweet* foi manualmente rotulado pelo autor deste trabalho como “positivo”, “neutro” ou “negativo”.

Com o objetivo de construir uma base de dados rotulada consistente, como exposto na Subseção 2.3.2, foi selecionada uma quantidade significativa de *tweets* para cada um dos filmes em questão, além de quantidades consideráveis de *tweets* que representam todas as semanas analisadas [3]. Também teve-se o cuidado de incluir quantidades significativas de *tweets* representando cada um dos três sentimentos, mesmo que realizar essa seleção seja um dos maiores desafios na construção de uma base de dados rotulada.

A Tabela 2 sumariza os dados contidos na base de *tweets* rotulados, que representa 0,36% da nova base após o pré-processamento.

Sentimento	Quantidade de *tweets*
positivo	1444
negativo	1362
neutro	429
Total	3235

Tabela 2 – Sumarização dos dados contidos na base de *tweets* rotulados

[3] A base rotulada utilizada neste estudo está disponível em http://bit.ly/TCCIgor

Em seguida, a base de dados rotulada foi submetida às etapas de pré-processamento descritas na Seção 3.3.

3.3 Pré-processamento de *tweets*

Uma vez que a coleta dos *tweets* é realizada, os dados precisam ser tratados, com o propósito de descartar o que é irrelevante para a etapa de classificação (FELIX, 2016). As etapas de pré-processamento realizadas neste trabalho são listadas adiante[4].

- Conversão de letras maiúsculas em minúsculas, com o objetivo de padronizar o texto.
- Remoção de *links*, uma vez que esses termos não possuem conteúdo semântico.
- Substituição de *emoticons* por palavras correspondentes. Nas redes sociais, frequentemente usuários utilizam *emoticons* para expressar alguma emoção. Um dicionário incluindo 136 *emoticons* ocidentais e suas respectivas traduções para a língua inglesa foi construído para este estudo.

  O dicionário foi construído a partir de experiência do autor e da lista de *emoticons* disponível na Wikipedia[5]. Na Tabela 3 são mostrados alguns exemplos de *emoticons* que foram traduzidos durante o pré-processamento.

Sentimentos positivos		Sentimentos negativos	
Emoticon	Tradução	*Emoticon*	Tradução
:D	laughing	:'(	crying
:-)	happy	:/	disappointed
=P	playful	:@	angry
<3	love	:X	uncomfortable

Tabela 3 – Exemplos de *emoticons* e suas respectivas traduções para a língua inglesa

- Remoção de caracteres não alfabéticos e pontuação, pois não agregam valor à classificação.
- Remoção de citação a outros usuários (no *Twitter*, precedidos de "@").
- Remoção dos títulos dos filmes. Os nomes dos filmes em questão foram removidos, uma vez que a presença desses termos poderia equivocar os resultados gerados pelos classificadores. Por exemplo: a palavra "hell", presente no título do filme *Hell or High Water*, geralmente expressa um sentimento negativo, o que poderia corromper a classificação de um *tweet* com sentimento positivo.

[4] A base original e a pré-processada utilizadas neste estudo estão disponíveis em http://bit.ly/TCCIgor
[5] http://bit.ly/wikiEmot

- Remoção de letras repetidas. Como mostrado na Figura 5, é comum que usuários das redes sociais repitam letras de palavras para intensificar o sentimento. Porém, essas palavras com letras repetidas não são reconhecidas pelos classificadores, portanto, as letras repetidas foram removidas.

Tweet original	*Tweet* após remoção de letras repetidas
I'm a mess after seeing Manchester By The Sea . Tears cried. Pants shat. MOVIE OF THE YEAAAAAAARRRR.	I'm a mess after seeing Manchester By The Sea . Tears cried. Pants shat. MOVIE OF THE YEAR.
UGHHHHHH i cant believe i wasted two hours watching fences	UGH i cant believe i wasted two hours watching fences

Tabela 4 – Exemplos de *tweets* após pré-processamento para remoção de letras repetidas

- Substituição de gírias e abreviações por expressões completas. Usuários do *Twitter* frequentemente utilizam gírias com o intuito de economizar palavras, já que os *tweets* são limitados a 140 caracteres. Um dicionário contendo 367 gírias e suas respectivas traduções foi construído para incorporar novos termos aos *tweets* e garantir que a semântica do *tweet* fosse preservada.

  O dicionário foi construído a partir de experiência do autor e de listas de gírias disponíveis na *Internet*[6] [7] [8]. Alguns dos termos presentes no dicionário, construído para este estudo, são apresentados na Tabela 5.

Gíria	Tradução	Gíria	Tradução
aight	all right	pls	please
ftw	for the win	h8	hate
omg	oh my God	thx	thanks
luv	love	zzz	boring

Tabela 5 – Exemplos de gírias e suas respectivas traduções para a língua inglesa

- Remoção de *stop words*. A Tabela 6 mostra exemplos de *stop words*, palavras que são bastante comuns em um idioma e, portanto, não possuem muito valor semântico. Por isso, são removidas durante o pré-processamento (FILHO, 2014). Uma lista com 569 *stop words*, parte do *Onix Text Retrieval Toolkit*[9], foi utilizada nesta etapa do pré-processamento.

- Remoção de *tweets* não relacionados aos filmes. Um dos desafios em Análise de Sentimentos é garantir que os dados sendo analisados correspondem ao tema em

[6] http://www.illumasolutions.com/omg-plz-lol-idk-idc-btw-brb-jk.htm
[7] http://allusefulinfo.com/whats-the-full-form-of-lol-asap-tos-btw-brb-other-50-most-common-abbreviations/
[8] https://kb.iu.edu/d/adkc
[9] http://www.lextek.com/manuals/onix/stopwords1.html

a	all	an	and
at	as	be	but
do	even	from	go
hello	just	look	too
thus	who	with	you

Tabela 6 – Exemplos de *stop words*

questão. Os títulos dos filmes *Arrival*, *Fences*, *La La Land*, *Lion* e *Moonlight* remetem a outras palavras e expressões utilizadas no cotidiano de pessoas que falam a língua inglesa, como pode ser visto nas Figuras 3 e 10.

Figura 10 – Na busca pelo filme *Fences*, alguns *tweets* referenciavam cercas e o muro que Donald Trump, presidente dos Estados Unidos, pretende construir para dificultar a entrada ilegal de imigrantes no país.

Além disso, os títulos dos filmes em questão também podem remeter a outros filmes, músicas, etc. Isso pode ser percebido em alguns *tweets*, como mostrado nas Figuras 11 e 12.

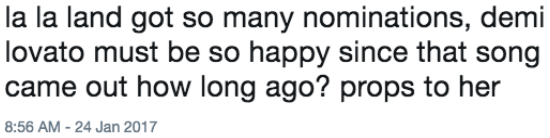

Figura 11 – *Tweet* coletado em consulta pelo filme *La La Land* que faz uma brincadeira com a música de mesmo nome da cantora Demi Lovato.

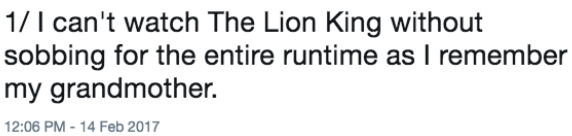

Figura 12 – Alguns *tweets* coletados na busca pelo filme *Lion* remetem ao filme da Disney, *The Lion King*.

Com o objetivo de minimizar os efeitos desse desafio e garantir que os *tweets* sendo classificados realmente fazem referência aos filmes indicados ao *Oscar 2017*, para cada um desses cinco filmes foi criada uma lista de termos não relacionados, exemplificada na Tabela 7. Os *tweets* contendo pelo menos um dos termos presentes na lista foram removidos da base.

Termos não relacionados			
Arrival		*Fences*	
new arrival	plane	picket	neighbor
aircraft	airport	wall	refugee
dead on arrival	flight	trump	border

Tabela 7 – Algumas palavras presentes na lista de termos não relacionados aos filmes *Arrival* e *Fences*. Todos os *tweets* contendo ao menos um termo da lista foram removidos da base por não terem relação com os respectivos filmes.

Para este trabalho, foi construído um algoritmo na linguagem *Java*[10] que realiza todas as etapas de pré-processamento listadas nesta Seção. O procedimento foi realizado individualmente para a base de dados de cada filme.

A quantidade de *tweets* que compõem a base de dados diminuiu devido à remoção dos *tweets* não relacionados aos filmes em questão e também daqueles que continham apenas termos irrelevantes à classificação (*tweets* contendo apenas *links* ou apenas *stop words*, por exemplo). A Tabela 8 mostra uma comparação da quantidade de *tweets* presentes na base antes do pré-processamento e após a realização dessa etapa – a última coluna indica quanto a base de *tweets* para cada filme decresceu após o pré-processamento.

Filme	Quantidade de *tweets*		
	Antes do pré-processamento	Depois do pré-processamento	Diferença
Arrival	138.825	135.214	-2,6%
Fences	53.211	41.682	-21,7%
Hacksaw Ridge	54.689	48.740	-10,9%
Hell or High Water	14.919	13.320	-10,7%
Hidden Figures	145.868	137.151	-6%
La La Land	250.942	244.213	-2,7%
Lion	186.295	150.641	-19,1%
Manchester by the Sea	31.768	28.601	-10%
Moonlight	108.121	90.278	-16,5%
Total	1.035.739	889.840	-14%

Tabela 8 – A quantidade original de *tweets* coletados e a quantidade de *tweets* que restou após realização do pré-processamento, indicando os dados que compõem a nova base de experimentos.

3.4 Classificação de *tweets*

Nesta etapa, são descritos algoritmos de classificação da literatura que foram utilizados para classificar cada *tweet* como "positivo", "neutro" ou "negativo". A forma como cada um dos algoritmos foi abordado é explicada nas subseções a seguir.

[10] O algoritmo desenvolvido para estudo está disponível em http://bit.ly/TCCIgor

Para este estudo, foram considerados três diferentes diferentes abordagens: aprendizado supervisionado, aprendizado por supervisão à distância e função de polaridade.

3.4.1 Algoritmo de aprendizado supervisionado

O algoritmo de aprendizado supervisionado *Naive Bayes* multinomial foi escolhido para ser usado neste trabalho por ser bastante utilizado em tarefas de classificação de textos e análise de sentimentos e apresentar bons resultados. Para aplicá-lo na base de dados rotulada, foi utilizado o *Weka* (SMITH; FRANK, 2016), um *software* escrito na linguagem *Java* que fornece uma coleção de algoritmos de aprendizagem de máquina para realização de tarefas relacionadas à mineração de dados. Além disso, ferramentas para pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização estão disponíveis no *Weka*.

Primeiramente, os dados da base rotulada de *tweets* foram convertidos para o formato *ARFF – Attribute Relation File Format –*, um arquivo de texto específico para o *Weka*, que descreve uma lista de instâncias que compartilham os mesmos atributos (PAYNTER et al., 2002). Dessa forma, é possível utilizar a base de dados rotulada como conjunto de treinamento para o algoritmo *Naive Bayes*.

As primeiras linhas do arquivo *ARFF* utilizado para este trabalho podem ser vistas na Figura 13. Os dois atributos da relação *OSCAR* são descritos adiante.

- *tweet*: uma *string* contendo o texto do *tweet*
- *class*: o rótulo assinalado manualmente para o *tweet* (positivo, neutro ou negativo)

```
@relation OSCAR

@attribute tweet string
@attribute class {pos,neg,neu}

@data
'nocturnal animals easily best movies',pos
'rethinking awesome',pos
'crazy good movie',pos
'best picture year locks',pos
'watched night interstellar vibes worth watch',pos
'good shit',pos
'best picture nomination predictions jackie silence',pos
'favorite movie bvs choose',pos
'blu ray',neu
'theaters',neu
'shut care',neu
'movie night started stand',neu
'watch mind fuckin blown',pos
'interstellar best movies watched',pos
```

Figura 13 – Trecho do arquivo *ARFF* utilizado para alimentar o *Weka*

Após carregar o conjunto de treinamento na interface *Explorer* do *Weka*, uma visão geral dos dados é mostrada na aba de pré-processamento, como pode ser observado na Figura 14.

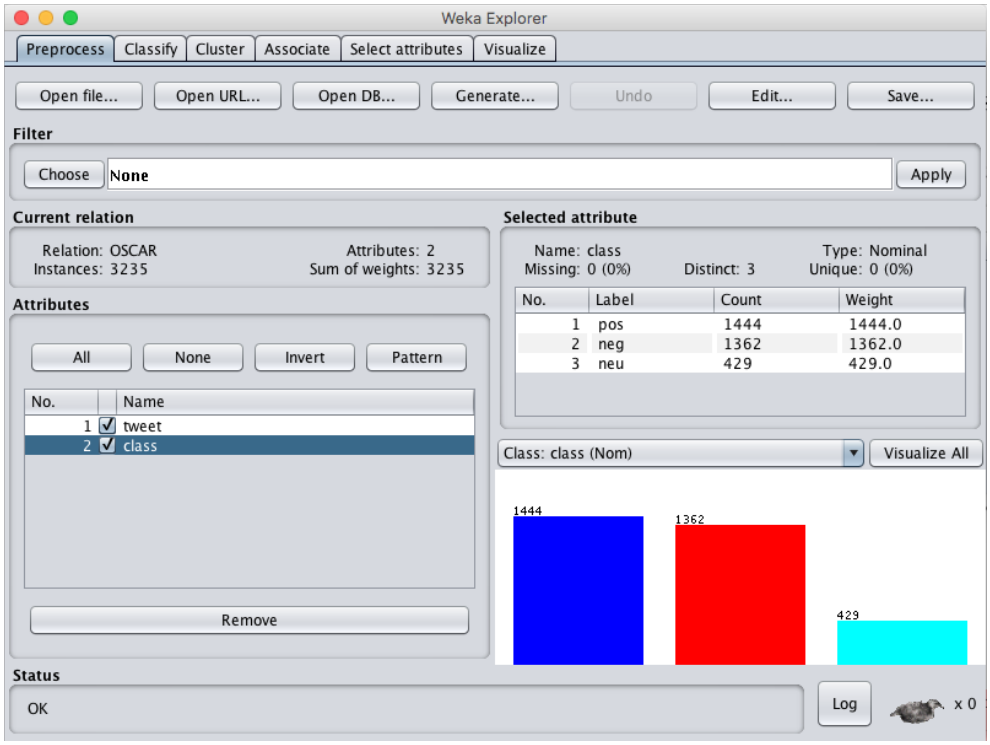

Figura 14 – Visão geral do conjunto de treinamento carregado no *Weka*

Em seguida, na aba de classificação, são escolhidos o classificador *Naive Bayes* multinomial e a estratégia de validação *10-fold cross-validation*, que consiste na divisão do conjunto de teste em 10 partes aleatórias, que são testadas individualmente, utilizando as outras nove partes restantes como conjunto de treinamento (GIGLIOTTI, 2012).

Foram utilizadas as configurações padrão do *Weka* para o *Naive Bayes* multinomial.

3.4.2 Algoritmo de aprendizado por supervisão à distância

Sentiment140 é uma ferramenta para análise de sentimentos específica para o *Twitter*. A classificação de *tweets* utilizando esse método é bastante simples, uma vez que há uma API disponível *online*[11]. Além disso, por ser um algoritmo de aprendizado por supervisão à distância, o modelo já foi treinado, então basta usar o modelo pronto – disponibilizado pela API – para classificar os *tweets*.

Para submeter vários *tweets* ao classificador, basta executar o comando especificado na Figura 15 por meio do Terminal, de modo a acessar a API. Assim, a classificação em massa é realizada.

A saída do classificador inclui o valor de polaridade atribuída e o seu respectivo *tweet*, como mostrado na Figura 16. Um *tweet* com polaridade de valor "0" indica sen-

[11] http://help.sentiment140.com/api

```
curl --data-binary @SENTIMENT.txt "http://www.sentiment140.com/api/bulkClassify"
```

Figura 15 – Submissão do arquivo de texto *SENTIMENT.txt*, em que cada linha contém um *tweet*, ao classificador *Sentiment140*

timento negativo, enquanto "2" designa neutralidade e "4" classifica uma sentença como positiva.

```
"4","film tht makes feel good strange life life film magical moments happy"
"2","funny times laughing loud"
"2","lit"
"2","casey affleck brilliant"
"4","hey super late great movie"
"2","brilliant"
"0","sad movie funny life"
"2","depressing film beautiful raw heartbreaking huhu"
"2","beautiful tragic calm watching"
"2","interesting film"
"2","night watched pleasantly surprised"
"2","watched fantastic film"
"2","hours thinking sign great movie"
"4","good movie great movie movie year"
"4","rated awesome imdb"
"2","yesterday depressing good"
"2","great movie"
"2","masterpiece"
"2","good job cassey affleck"
"2","masterfully told beautifully acted shattering graceful elegy loss grief friday"
"2","fantastic characters written beautifully"
"2","best films years"
"2","finally exceptional"
"2","good movie heartbreaking pick"
"2","casey affleck lucas hedges brilliant"
"2","pretty amazing"
"2","recommend great"
"2","watched film"
"2","great movie accurate depiction lives unravel coach eric taylor died"
"2","fucking good"
"2","front runner movie year top"
"4","rated awesome imdb"
```

Figura 16 – Trecho do retorno da classificação realizada com o *Sentiment140*

Para este estudo, o comando citado na Figura 15 foi executado três vezes, uma para cada um dos arquivos individuais de *tweets* que foram previamente separados por sentimentos. Dessa forma, sabe-se a quantidade de *tweets* que foi classificada pelo *Sentiment140* com o sentimento correto.

3.4.3 Função de polaridade

Neste trabalho, a função *polarity*, da biblioteca *TextBlob*, também foi avaliada em relação à classificação de *tweets*. Para se classificar *tweets* em massa, foi necessário escrever um simples código na linguagem *Python*, exposto na Figura 17.

Basicamente, após a importação da biblioteca *TextBlob*, cada uma das linhas do arquivo *TEXTBLOB.csv* é percorrida por meio de um laço e transformada em uma estrutura de dados chamada *TextBlob*, aceita pela função, possibilitando a realização do cálculo da polaridade do texto. Apenas um *tweet* está presente em cada linha do arquivo.

```
>>> from textblob import TextBlob
>>> with open("TEXTBLOB.csv") as f:
...     for line in f:
...             testimonial = TextBlob(line)
...             testimonial.sentiment.polarity
```

Figura 17 – Código em *Python* para percorrer um arquivo de *tweets* e classificar utilizando a biblioteca *TextBlob*

Após a execução do código, são retornados os valores de polaridade para cada um dos *tweets* presentes no arquivo, como mostrado na Figura 18.

```
0.24999999999999997
0.0
0.0
0.05000000000000001
0.0
0.375
0.8
0.3
0.85
-0.07499999999999997
0.20000000000000004
0.9
0.28
0.675
0.6
0.85
0.6000000000000001
0.23333333333333336
0.08642857142857141
0.2833333333333334
0.35
0.7
0.0
0.4166666666666667
0.6000000000000001
-0.4
0.575
0.175
0.44999999999999996
1.0
0.2
0.15
0.4166666666666667
0.0
```

Figura 18 – Parte dos valores de polaridade obtidos com a classificação utilizando o *TextBlob*

Os valores de polaridade obtidos, exemplificados na Figura 18, são números entre "-1" e "1". Valores de polaridade mais próximos de "1" sugerem maior intensidade de sentimento positivo para a sentença, enquanto que valores mais próximos de "-1", apontam para maior intensidade de sentimento negativo na sentença; os valores próximos de "0" indicam neutralidade.

Para este trabalho, o código da Figura 17 foi rodado três vezes, ou seja, uma para cada um dos arquivos individuais de *tweets*, previamente separados por sentimentos. Consequentemente, sabe-se a quantidade de *tweets* que foi classificada corretamente pelo *TextBlob*.

3.5 Avaliação dos classificadores

Cada um dos classificadores foi testado utilizando a base de dados rotulada e a estratégia de amostragem usada foi a validação cruzada de 10 pastas, com a medida das 10 execuções sendo apresentada.

A acurácia foi utilizada neste trabalho para validar os classificadores. Sabe-se que essa medida tem algumas deficiências (TAN; STEINBACH; KUMAR, 2005), como a possibilidade de não ser adequada para classes desbalanceadas, pois tende a privilegiar a classe majoritária – sendo que, normalmente, a classe rara é mais interessante. Porém, o problema em questão apresenta classes pouco desbalanceadas e, por questões de simplicidade e por não ser o foco deste trabalho, somente essa medida foi utilizada para realizar a validação dos classificadores.

3.6 Criação de *ranking* do *Oscar 2017*

Não existe um *ranking* que possa classificar os filmes que concorreram ao *Oscar 2017* de acordo com seu desempenho na premiação, apenas são disponibilizadas as informações sobre a quantidade de indicações e vitórias que cada filme acumulou na premiação.

Dessa forma, para facilitar a comparação do resultado do *Oscar 2017* com aquele obtido após a classificação, decidiu-se criar para este estudo uma medida para obter um *ranking* dos filmes em questão.

Ao visitar o *website* oficial da premiação[12], percebe-se que a representação visual dos vencedores do *Oscar* é apresentada em três tamanhos diferentes. Na página, as categorias de maior relevância têm mais destaque (Figura 19) e, portanto, podem ser consideradas as que têm maior peso na premiação, como Melhor Filme e Melhor Animação.

Assim, para compor a pontuação de cada filme e posteriormente construir o *ranking*, decidiu-se que as categorias com maior destaque têm peso "3", as com destaque médio têm peso "2", e, por fim, as categorias com menor destaque têm peso "1". Foram consideradas apenas as categorias do *Oscar 2017* em que pelo menos um dos indicados a Melhor Filme está presente.

Assim sendo, montou-se a Tabela 9, que mostra as categorias em questão e os respectivo peso (p) que cada uma representa na pontuação do filme. Apesar de a categoria Melhor Direção ter menor destaque no *website* da premiação, para este trabalho ela foi considerada com peso médio, por ter mais relevância que as outras categorias consideradas de peso baixo.

As quantidades de vitórias e indicações conquistadas por cada filme no *Oscar 2017*

[12] http://oscar.go.com/winners

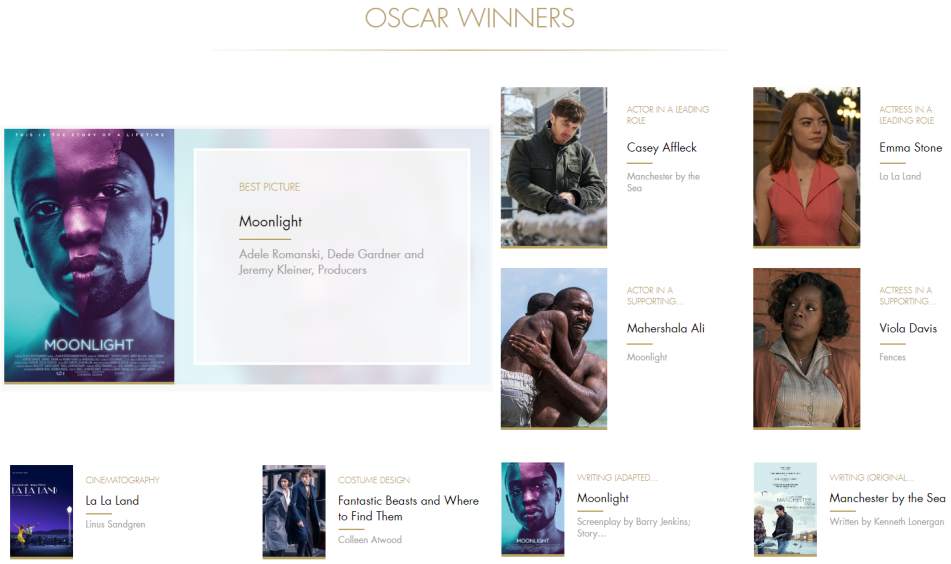

Figura 19 – Alguns dos vencedores do *Oscar 2017* e a forma como são dispostos no *website* oficial da premiação. As categorias com maior destaque são consideradas mais relevantes. – Fonte: 89$^{\text{th}}$ Academy Awards (http://oscar.go.com/winners)

n	Categorias	p
1	Filme	3
2–6	Diretor, Ator, Atriz, Ator Coadjuvante, Atriz Coadjuvante	2
7–16	Roteiro Original, Roteiro Adaptado, Trilha Sonora, Canção Original, Edição de Som, Mixagem de Som, Produção, Cinematografia, Figurino, Edição	1

Tabela 9 – Os pesos que cada uma das categorias em questão possui.

foram obtidas a partir do *website* oficial da premiação. Uma vez em posse desses dados e dos pesos indicados na Tabela 9, a pontuação de cada filme pode ser calculada de acordo com a Equação 3.1.

Nessa Equação, n indica o índice, representado na Tabela 9, de cada uma das 16 categorias consideradas. Além disso, o peso a ser multiplicado pelo número de indicações recebidas pelo filme é reduzido à metade, já que possui menor importância que as vitórias.

$$pontos_{\text{filme}} = \sum_{n=1}^{16} (\#vit_{\text{n}} \times p_{\text{n}}) + (\#ind_{\text{n}} \times \frac{p_{\text{n}}}{2}) \tag{3.1}$$

3.7 Ferramentas para comparar o resultado do *Oscar 2017* com os sentimentos dos *tweets*

Uma vez criado o *ranking* do *Oscar 2017*, diferentes abordagens foram usadas para tentar estabelecer uma relação entre esse *ranking* e o sentimento obtido a partir da classificação dos *tweets*.

- Criação de *rankings* de acordo com os sentimentos dos *tweets*: diferentes *rankings* foram construídos a partir dos resultados obtidos com o classificador escolhido e posteriormente foram analisados e comparados em relação ao resultado do *Oscar 2017*.

  Alguns dos *rankings* construídos foram baseados na quantidade de *tweets* que foi recuperada para cada filme em questão, na quantidade de *tweets* classificados com sentimento positivo (os filmes com maior índice de instâncias positivas ocupam posições maiores), entre outros.

- Coeficiente de correlação de postos de *Spearman*: essa medida calcula a intensidade da relação entre variáveis ordinais utilizando apenas a ordem das observações[13], portanto, pode calcular a intensidade da relação entre dois *rankings*. Quanto mais próximo de -1, mais forte e negativa é a associação entre as variáveis. Quanto mais próximo de 1, mais forte e positiva é a associação entre as variáveis.

- Análise de gráficos: gráficos incluindo os dados obtidos a partir da coleta de *tweets* e da classificação serão construídos e analisados.

- Nuvem de palavras: foram construídas representações gráficas da frequência de ocorrência dos termos contidos nos *tweets* referentes a cada filme. As palavras com maior destaque indicam maior ocorrência nos *tweets*.

- Gráficos: foram construídos para facilitar a visualização e análise dos dados.

3.8 Considerações finais

Durante o desenvolvimento deste estudo, as etapas da tarefa de Análise de Sentimentos foram aplicadas a um problema prático para extrair informação útil a partir de dados do *Twitter*. Na primeira etapa, os *tweets* relacionados aos filmes indicados à categoria de Melhor Filme do Oscar 2017 foram coletados e pré-processados.

Em seguida, uma amostra aleatória da base de *tweets* original foi separada com o objetivo de construir uma base de dados rotulada para ser utilizada em experimentos que

[13] http://bit.ly/SpearmanESTGV

envolvem modelo de aprendizagem supervisionado. A cada *tweet* que compõe essa base foi manualmente atribuído um rótulo de sentimento – positivo, neutro ou negativo.

Então, foi possível testar três diferentes classificadores. Cada um foi avaliado e o que obteve o melhor desempenho foi escolhido para classificar os *tweets* da base de dados completa. Assim, foi possível predizer o sentimento predominante dos usuários do *Twitter* com relação aos filmes do *Oscar 2017* e criar um *ranking* de acordo com a preferência dos usuários da rede social.

Para comparar o resultado do classificador com o do *Oscar 2017*, foi desenvolvida uma medida para criar um *ranking* da premiação, especialmente para este estudo. Os resultados, então, foram comparados a partir do coeficiente de correlação de postos de *Spearman* e da análise de gráficos.

4 Resultados

Este capítulo tem como objetivo analisar os dados obtidos pelo pré-processamento dos *tweets* e sua relação com o resultado do *Oscar 2017*. Para isso, diferentes comparações foram realizadas.

A Seção 4.1 apresenta uma visão geral do *Oscar 2017*, incluindo a quantidade de indicações e vitórias que cada filme conquistou nas categorias da premiação. O *ranking* do *Oscar 2017*, criado para este trabalho, também é exposto.

A Seção 4.2 sumariza a quantidade de *tweets* coletados, a composição da base e tenta estabelecer relações entre esses resultados e o do *Oscar 2017*. Observações são realizadas a partir da análise de gráficos que resumem esses dados.

A Seção 4.3 tem como objetivo mostrar o resultado da avaliação dos classificadores a partir da análise de matrizes de confusão e do cálculo da acurácia para cada classificador analisado.

A Seção 4.4 apresenta os resultados da classificação dos *tweets* com o classificador escolhido, *Naive Bayes*.

A Seção 4.5 apresenta nuvens de palavras referentes aos *tweets* de alguns filmes e discute sobre a qualidade do pré-processamento realizado na base.

A Seção 4.6 apresenta a comparação do sentimento dos *tweets* em relação ao *Oscar 2017*, incluindo uma análise minuciosa de tabelas e gráficos construídos com base nos resultados obtidos a partir do classificador.

4.1 Visão geral do *Oscar 2017*

Os indicados às categorias do *Oscar 2017* foram apresentados ao público em 24 de Janeiro de 2017 (DONNELLY, 2017). A 89ª edição do *Oscar*, apresentada pela Academia de Artes e Ciências Cinematográficas, aconteceu no dia 26 de Fevereiro de 2017 em Los Angeles, Estados Unidos, onde os vencedores foram anunciados (DONNELLY, 2017).

Para este estudo, levou-se em conta apenas o desempenho dos filmes indicados à categoria de Melhor Filme nessa edição. A Tabela 10 mostra quantas indicações e vitórias cada um dos filmes em questão acumulou dentre todas as categorias do *Oscar 2017*.

Para o *Oscar 2017*, a Academia premiou 24 categorias. Dentre essas, os filmes analisados estavam envolvidos em 16 e, no total, apenas uma categoria dessas categorias não foi vencida por um dos filmes em questão.

Filme	Indicações		Vitórias	
	Quantidade	%	Quantidade	%
Arrival	8	14%	1	7%
Fences	4	7%	1	7%
Hacksaw Ridge	6	10%	2	13%
Hell or High Water	4	7%	0	0%
Hidden Figures	3	5%	0	0%
La La Land	14	24%	6	40%
Lion	6	10%	0	0%
Manchester by the Sea	6	10%	2	13%
Moonlight	8	14%	3	20%
Total	59	100%	15	100%

Tabela 10 – Relação das indicações e vitórias acumuladas por cada um dos filmes indicados à categoria de Melhor Filme do *Oscar 2017*

A Figura 20 mostra por meio de um gráfico de barras os dados da Tabela 10, onde é possível perceber que os líderes de indicações e vitórias são os filmes *La La Land* e *Moonlight*. Também é possível constatar que os filmes *Hell or High Water*, *Hidden Figures* e *Lion* não venceram em nenhuma categoria.

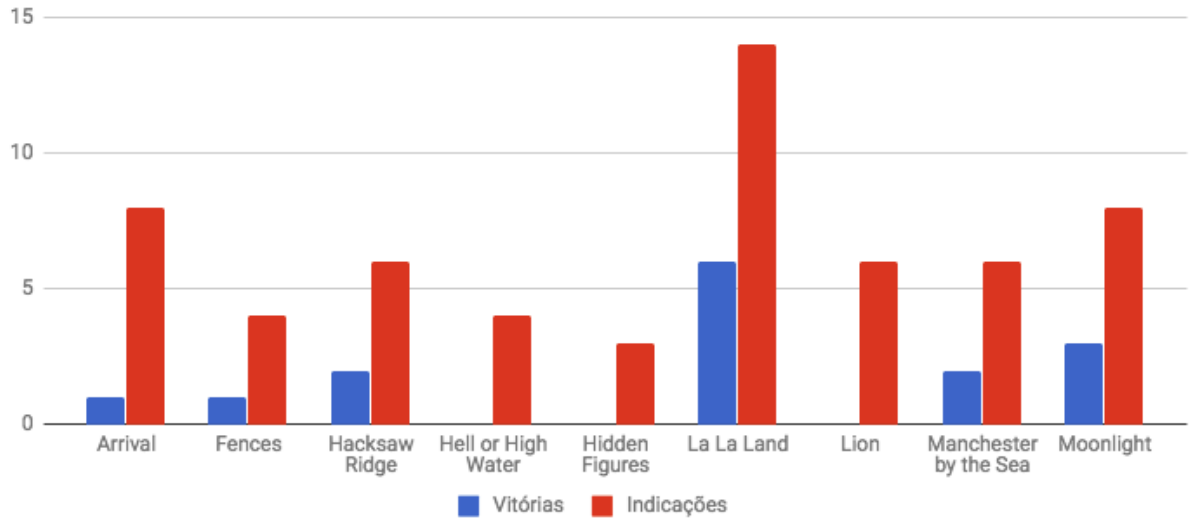

Figura 20 – Indicações e vitórias acumuladas por cada filme em questão

Uma vez em posse desses dados, foi possível construir o *ranking* dos filmes indicados à categoria de Melhor Filme do *Oscar 2017* em conformidade com as pontuações obtidas de acordo com a medida discutida na Seção 3.6. O *ranking* pode ser observado na Tabela 11.

Posição	Filme	Pontuação
1	*La La Land*	17,5
2	*Moonlight*	12,5
3	*Manchester by the Sea*	9
4	*Hacksaw Ridge*	7
5	*Arrival*	6,5
6	*Fences*	6
7	*Lion*	5
8	*Hell or High Water*	3,5
9	*Hidden Figures*	3

Tabela 11 – *Ranking* dos filmes indicados à categoria de Melhor Filme do *Oscar 2017*, construído especialmente para este estudo

4.2 Quantidade de *tweets*

A Tabela **??**, já discutida no Capítulo 3, mostra a quantidade de *tweets* que foi recuperada a partir da coleta de dados. Devido à remoção de *tweets* não relacionados aos filmes, etapa de pré-processamento discutida na Seção 3.3, a quantidade de *tweets* que compõem a base de estudos diminuiu, como pode ser visto na Tabela 8.

A Figura 21 mostra a quantidade de *tweets* presentes na base para cada filme de acordo com a semana. Pode-se perceber que a maior ocorrência de *tweets* para todos os filmes aconteceu durante a primeira semana de análise, isto é, na semana em que os indicados ao *Oscar 2017* foram revelados.

A quantidade de *tweets* para todos os filmes diminuiu drasticamente durante a segunda semana e, ao longo da terceira semana, aumentou para os filmes *Fences*, *Hacksaw Ridge*, *Hell or High Water*, *Manchester by the Sea* e *Moonlight*. Dentre todos os filmes, apenas *Fences*, *Hacksaw Ridge* e *Lion* tiveram uma diminuição na ocorrência de *tweets* durante a quarta semana.

Exceto por *Arrival*, *Fences* e *Hidden Figures*, todos os outros filmes tiveram um aumento na ocorrência de *tweets* durante a última semana de análise, que finalizou um dia antes do *Oscar 2017*.

É possível perceber uma relação entre a movimentação na rede social e o *Oscar 2017*, já que a quantidade de *tweets* publicados sobre os filmes na semana em que os indicados foram anunciados foi maior do que nas outras. Além disso, mais próximo da data em que ocorreu a premiação, a quantidade de *tweets* relacionados à maioria dos filmes aumentou gradativamente.

Com base na Figura 21, pode-se perceber que os filmes *Arrival*, *Hidden Figures*, *La La Land*, *Lion* e *Moonlight* foram mais comentados do que os outros filmes durante o período de análise. Isso pode ser explicado pelo fato de que esses filmes estrearam em

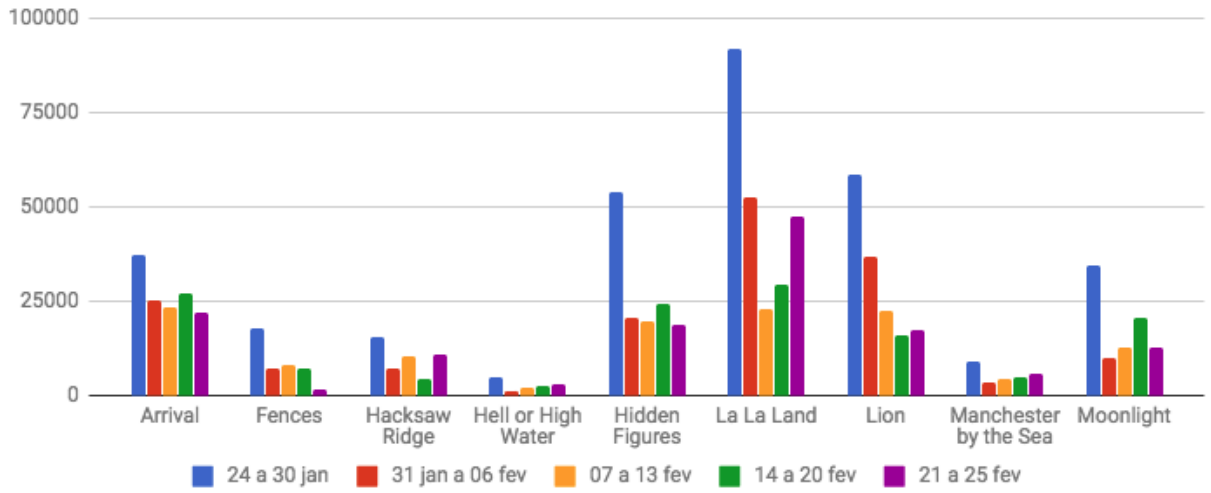

Figura 21 – Quantidade de *tweets* publicados para cada filme por semana

cinemas dos Estados Unidos entre Novembro de 2016 e Janeiro de 2017[1], época mais próxima da cerimônia de entrega dos *Oscars*.

Além disso, outra explicação pode ser feita a partir do fato de que *Arrival*, *La La Land* e *Moonlight* são filmes que figuram entre as cinco primeiras posições do *ranking* do *Oscar 2017*. Também vale salientar que esses três filmes foram os que mais receberam indicações à premiação e os dois últimos foram os que mais ganharam prêmios.

Apesar de *Hacksaw Ridge* ter estreado em Novembro de 2016, *Fences* em Dezembro de 2016[2] e ambos ocuparem, respectivamente, a quarta e sexta posições do *ranking* do *Oscar 2017*, esses dois filmes não receberam tanta atenção dos usuários do *Twitter* em comparação aos cinco filmes citados anteriormente. Isso pode ter acontecido devido à falta de divulgação dos longa-metragens ou simplesmente for falta de interesse do público.

Os dois filmes menos comentados durante o período de análise, *Hell or High Water* e *Manchester by the Sea*, podem ter assumido essas posições por terem estreado mais cedo. O primeiro foi lançado nos cinemas dos Estados Unidos em Agosto de 2016, enquanto o segundo recebeu mais atenção do público ao estrear no famoso *Sundance Film Festival* (Estados Unidos), em Janeiro de 2016[3].

Ao analisar o gráfico presente na Figura 22, também é interessante perceber que a euforia dos usuários do *Twitter* foi maior durante a primeira semana de análise, em que os filmes indicados foram divulgados. A quantidade de *tweets* diminuiu significamente durante as próximas duas semanas e teve um leve aumento na penúltima semana e na semana que antecedeu a premiação. Uma possível explicação pode ser devido à grande expectativa de fãs de cinema sobre quais filmes receberiam indicações ao *Oscar*, causando

[1] http://www.imdb.com
[2] Ver nota de rodapé 1
[3] Ver nota de rodapé 1

a publicação de mais *tweets* próximo à divulgação dos indicados. É provável que o número de *tweets* tenha aumentado logo após a cerimônia de entrega dos prêmios.

Também é possível perceber que o filme *La La Land* foi o mais comentado durante quatro semanas – apenas foi superado por *Arrival*, na terceira semana –, e isso corresponde com o fato de que esse filme ocupa o primeiro lugar no *ranking* do *Oscar 2017* (Tabela 11). Nota-se que o filme *Hell or High Water* foi o menos comentado durante todas as semanas, e isso também é algo que corresponde com o filme ocupando o penúltimo lugar no *ranking* do *Oscar 2017*. Dessa forma, pode-se notar uma relação entre o *ranking* do *Oscar 2017* e a movimentação na rede social, que muitas vezes cresce de acordo com a visibilidade que certo filme recebe ao ser indicado à premiação.

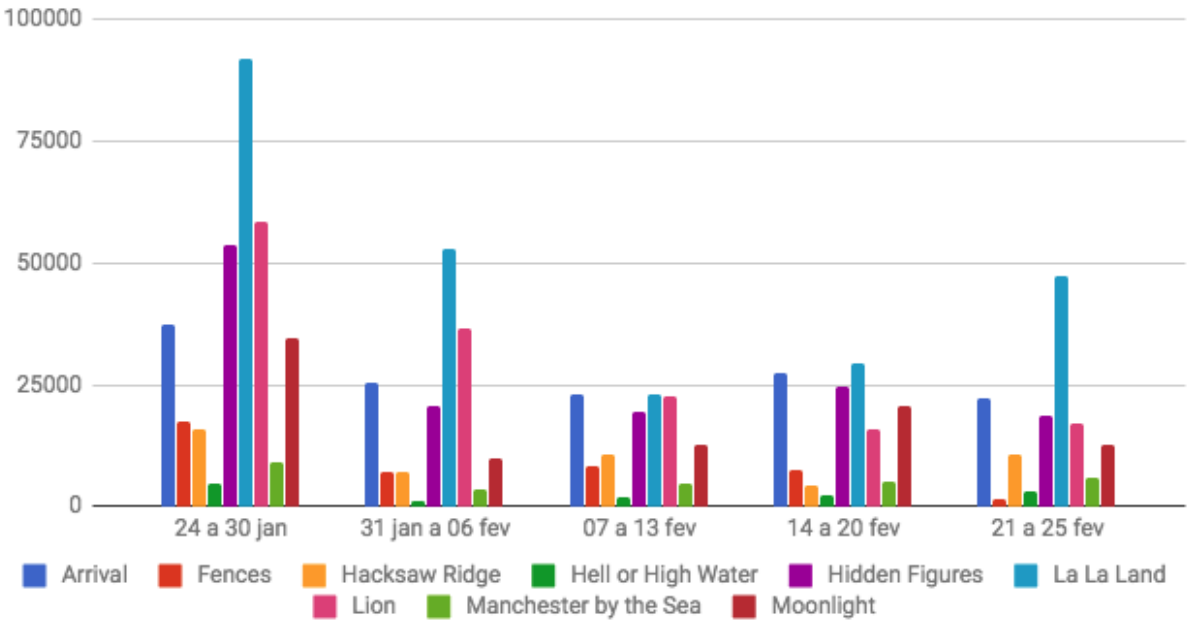

Figura 22 – Visão geral da quantidade de *tweets* publicados por semana para cada filme

4.3 Avaliação dos classificadores

Na subseção 2.3.5 foram discutidas medidas para avaliação do desempenho de classificadores, como a geração de matrizes de confusão e o cálculo da acurácia. Com o objetivo de escolher o melhor classificador dentre os três que este trabalho tem como foco, essas medidas foram levadas em consideração.

As Tabelas 12, 13 e 14 mostram a matriz de confusão para cada algoritmo de classificação utilizado neste trabalho. As medidas **VP** (verdadeiros positivos), **VN** (verdadeiros negativos) e **VE** (verdadeiros neutros) estão indicadas em negrito.

A Tabela 12, obtida a partir da geração do modelo *Naive Bayes* multinomial no *Weka*, mostra a matriz de confusão para tal algoritmo.

classificado como				
pos	neg	neu		
1191	158	95	pos	
279	**987**	96	neg	real
158	50	**221**	neu	
1628	1195	412		

Tabela 12 – Matriz de confusão para o algoritmo *Naive Bayes*

Como os classificadores Polaridade, do *TextBlob*, e *Sentiment140* foram aplicados individualmente para cada um dos três sentimentos, foi possível descobrir as medidas citadas na Seção 2.3.5 e construir as matrizes de confusão.

A matriz de confusão para o Polaridade, do *TextBlob*, é demonstrada na Tabela 13.

classificado como				
pos	neg	neu		
1081	80	283	pos	
364	**702**	296	neg	real
103	41	**285**	neu	
1548	823	864		

Tabela 13 – Matriz de confusão para o *TextBlob*

Para o *Sentiment140*, a matriz de confusão está representada na Tabela 14.

classificado como				
pos	neg	neu		
198	18	1228	pos	
51	**240**	1071	neg	real
16	3	**410**	neu	
265	261	2709		

Tabela 14 – Matriz de confusão para o *Sentiment140*

A partir da Equação 2.2 e dos dados das Tabelas 12, 13 e 14, construiu-se a Tabela 15, onde pode-se observar que o classificador *Naive Bayes* obteve maior acurácia dentre os avaliados. Portanto, esse será o algoritmo de classificação a ser utilizado para classificar a base completa de *tweets*.

Classificador	*Naive Bayes*	*TextBlob*	*Sentiment140*
Acurácia	74,1%	63,9%	26,2%

Tabela 15 – Acurácia calculada para cada um dos classificadores

4.4 Classificação dos *tweets* com o classificador escolhido – *Naive Bayes*

Após a geração do modelo de classificação baseado no algoritmo *Naive Bayes* pelo *Weka*, como mostrado na Subseção 3.4.1, é necessário fornecer os conjutos de teste – no caso, os arquivos individuais contendo todos os *tweets* coletados para cada um dos filmes indicados à categoria de Melhor Filme do *Oscar 2017* – para o classificador.

Esse processo foi realizado individualmente para cada um dos nove filmes em questão, resultando na seguinte Tabela 16, onde pode-se observar os sentimentos dos usuários do *Twitter* em relação aos filmes de acordo com classificação feita a partir do modelo *Naive Bayes* multinomial.

Filme	Sentimentos					
	Positivos		Negativos		Neutros	
Arrival	68.485	51%	27.002	20%	39.727	29%
Fences	20.199	48%	8.687	21%	12.796	31%
Hacksaw Ridge	31.998	66%	6.380	13%	10.362	21%
Hell or High Water	7.077	53%	2.064	15%	4.179	31%
Hidden Figures	77.314	56%	16.212	12%	43.625	32%
La La Land	104.507	43%	57.237	23%	82.469	34%
Lion	75.906	50%	36.978	25%	37.757	25%
Manchester by the Sea	14.879	52%	5.410	19%	8.312	29%
Moonlight	49.668	55%	17.130	19%	23.480	26%
Total	450.033	50,6%	177.100	19,9%	262.707	29,5%

Tabela 16 – Sentimentos dos usuários do *Twitter* em relação aos filmes de acordo com classificação feita a partir do modelo *Naive Bayes* multinomial

A partir da análise da Tabela 16, pode-se notar que, para todos os filmes, o número de *tweets* classificados como "positivo" é maior que o número dos classificados como "negativo" e "neutro". Esse fato também pode ser constatado claramente por meio do gráfico exposto na Figura 23. Isso é um indicativo de que os usuários escolheram passar mais tempo publicando *tweets* positivos sobre os filmes que gostaram a publicar críticas negativas sobre outros filmes que não lhes agradaram.

Há também um alto índice de *tweets* classificados como "neutro", o que acaba não agregando informações relevantes ao estudo. Sabe-se que o filme foi comentado, mas não se tem conhecimento sobre a natureza desse comentário, isto é, se o sentimento expresso no *tweet* é positivo ou negativo. Essa dificuldade acontece bastante na tarefa de Análise de Sentimentos e é um dos desafios da área.

É válido ressaltar que o perfil dos usuários do *Twitter* varia entre aqueles que são bastante empolgados, outros que fazem questão de publicar mensagens de ódio e

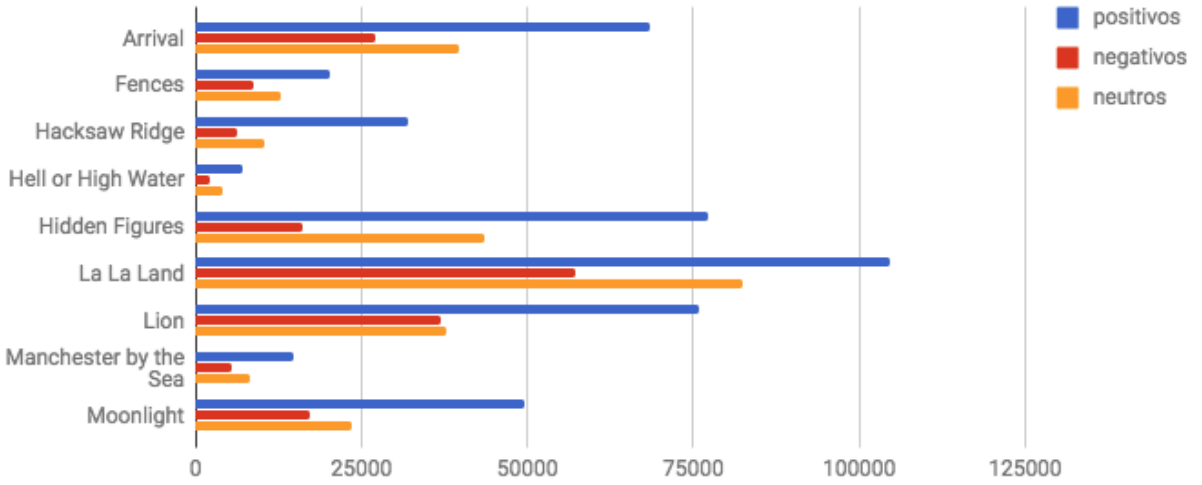

Figura 23 – Visualização gráfica dos sentimentos expressos pelos usuários do *Twitter* em relação aos filmes analisado

reclamações, entre outros. Dessa forma, o perfil do usuário pode influenciar na qualidade da informação, porém, neste estudo, todos os usuários foram tratados igualmente.

4.5 Validação do pré-processamento

Diversas etapas de pré-processamento, descritas na Seção 3.3, foram realizadas para tratar a base de *tweets* com o objetivo de descartar termos inúteis à classificação. Ao observar algumas nuvens de palavras, é possível perceber que o pré-processamento, em sua maioria, gerou bons resultados, já que a maior parte das palavras que aparecem nas nuvens faz referência aos filmes em questão e seus respectivos membros de elenco, equipe técnica, enredo, roteiro, entre outros.

Na Figura 24, que contém a nuvem de palavras referente aos *tweets* publicados sobre o filme *Hidden Figures*, pode-se perceber que palavras de natureza positiva, como "amazing", "fantastic", "inspirational" e "wonderful" foram bastante presentes entre os *tweets*. Alguns nomes de atores e personagens do filme são citados e vários termos relacionados a empoderamento de mulheres negras (por exemplo: "black women" e "blackhistorymonth") também aparecem, já que essa é uma questão social abordada pelo filme.

A nuvem de palavras para os *tweets* relacionados ao filme *La La Land* é exibida na Figura 25 e é possível perceber claramente que o público do *Twitter* falou bastante sobre a trilha sonora ("soundtrack" e "musical") do filme, que recebeu três indicações ao *Oscar 2017*, vencendo em duas categorias. Os nomes dos atores principais, Emma Stone – que venceu a categoria Melhor Atriz – e Ryan Gosling – que foi indicado à categoria Melhor Ator –, também aparecem na nuvem.

Todavia, a nuvem de palavras para o filme *Lion* apresenta algumas palavras não

Figura 24 – Nuvem de palavras contidas nos *tweets* coletados para o filme *Hidden Figures* — Gerada a partir do *WordItOut* (https://worditout.com/word-cloud/create)

Figura 25 – Nuvem de palavras contidas nos *tweets* coletados para o filme *La La Land* — Gerada a partir do *WordItOut* (https://worditout.com/word-cloud/create)

relacionadas ao filme, como “africa”, “captive” e “sheep”. Isso apresenta uma possível falha do pré-processamento, que poderia ter sido mais rigoroso na remoção de *tweets* não

relacionados a esse filme, isto é, mais palavras não-relacionadas a esse filme deveriam ter sido adicionadas ao filtro. Apesar disso, várias palavras relacionadas ao filme (por exemplo: "oscars", "dev patel" – o ator principal – e "watched") aparecem na nuvem.

Figura 26 – Nuvem de palavras contidas nos *tweets* coletados para o filme *Lion* — Gerada a partir do *WordItOut* (https://worditout.com/word-cloud/create)

Também é possível perceber que todas as palavras que aparecem nas nuvens estão completas, isto é, não apareceram gírias, nem abreviações, o que significa que os dicionários construídos para traduzir essas palavras funcionaram bem.

Comuns às nuvens de todos os filmes em questão estão palavras como "oscar", "best picture", "win" e "nowplaying", o que pode indicar que o *Oscar* é um motivador para discussões no *Twitter* e que seus usuários consideram os filmes indicados relevantes o bastante para serem comentados na rede social.

4.6 Comparação do sentimento dos *tweets* em relação ao *Oscar 2017*

O objetivo desta seção é comparar os sentimentos expressos pelos usuários do *Twitter* em relação aos filmes com o *ranking* do *Oscar 2017* (Tabela 11). Como pode ser observado na Figura 27, há vários indicadores que podem ser obtidos e transformados em *rankings*, possibilitando uma melhor análise dos dados.

No âmbito dos sentimentos expressos por meio do *Twitter*, obtidos a partir do resultado do classificador, três indicadores podem ser obtidos: a quantidade de *tweets* presentes na base para cada um dos filmes em questão, a quantidade de *tweets* positivos em

comparação com todas as instâncias da base classificadas como "positivo" e a quantidade de *tweets* positivos dentre todas as instâncias de certo filme classificadas como "positivo".

A partir do resultado do *Oscar 2017*, também derivam-se três indicadores: o número de indicações recebidas por cada filme, o número de vitórias conquistadas por cada um e o *ranking* criado para a premiação (Tabela 11).

Figura 27 – Indicadores que podem ser obtidos a partir dos dados disponíveis para este estudo. O objetivo é comparar os sentimentos expressos no *Twitter* em relação aos filmes com o resultado do *Oscar 2017*

A Tabela 17 mostra três *rankings* diferentes construídos com base nos indicadores obtidos a partir dos resultados do classificador. Eles foram construídos de acordo com o número total de *tweets* que compõem a base, o número de *tweets* positivos dentre os próprios de cada filme e o número de *tweets* positivos dentre todos os positivos da base.

Já a Tabela 18 mostra os três *rankings* que podem ser obtidos a partir do resultado do *Oscar 2017*. A segunda coluna reproduz o *ranking* do *Oscar 2017* construído para este estudo. Os outros dois *rankings* foram construídos de acordo com o número de indicações que cada filme recebeu e a quantidade de vitórias que cada um conquistou.

Filme	Base completa	Positivos (dentre os próprios)	Positivos (dentre todos os positivos)
Arrival	4º (15,2%)	6º (51%)	4º (15,2%)
Fences	7º (4,7%)	8º (48%)	7º (4,5%)
Hacksaw Ridge	6º (5,5%)	1º (66%)	6º (7,1%)
Hell or High Water	9º (1,5%)	4º (53%)	9º (1,6%)
Hidden Figures	3º (15,4%)	2º (56%)	2º (17,2%)
La La Land	1º (27,4%)	9º (43%)	1º (23,2%)
Lion	2º (16,9%)	7º (50%)	3º (16,9%)
Manchester by the Sea	8º (3,2%)	5º (52%)	8º (3,3%)
Moonlight	5º (10,1%)	3º (55%)	5º (11,0%)

Tabela 17 – *Rankings* construídos com base nos indicadores que podem ser obtidos a partir dos resultados dos sentimentos expressos no *Twitter* em relação aos filmes

Filme	*Ranking* do *Oscar 2017*	Indicações	Vitórias
La La Land	1º	1º (24%)	1º (40%)
Moonlight	2º	2º (14%)	2º (20%)
Manchester by the Sea	3º	4º (10%)	3º (13%)
Hacksaw Ridge	4º	4º (10%)	3º (13%)
Arrival	5º	2º (14%)	5º (7%)
Fences	6º	7º (7%)	5º (7%)
Lion	7º	4º (10%)	7º (0%)
Hell or High Water	8º	7º (7%)	7º (0%)
Hidden Figures	9º	9º (5%)	7º (0%)

Tabela 18 – *Rankings* construídos com base nos indicadores que podem ser obtidos a partir do resultado do *Oscar 2017*.

4.6.1 Análise intuitiva da correlação entre os dados

A partir da análise das Tabelas 17 e 18, pode-se perceber que o filme *La La Land* obteve a maior quantidade de *tweets* em relação à base completa, mas também teve a menor porcentagem de *tweets* classificados como "positivo" dentre os seus próprios *tweets*. Porém, *La La Land* obteve a maior quantidade de *tweets* classificados como "positivo" (Tabela 16) tanto em relação à base completa quanto em relação a todos os *tweets* classificados como "positivo" e isso pode indicar relação com o fato de que esse foi o filme que conquistou o maior número de indicações e vitórias, podendo considerá-lo o grande vencedor. A Tabela 18 também reforça que esse foi o filme ganhador do *Oscar 2017*, considerando a medida utilizada neste trabalho para obter o *ranking* da premiação.

Em contrapartida, nas Tabelas é mostrado que *Hell or High Water* pode ser considerado o perdedor se considerados esses dados. A quantidade de *tweets* relacionados a

esse filme representa apenas 1,5% da base de dados completa e é o filme que tem menos comentários positivos dentre todos os *tweets* classificados como "positivo", o que mostra que não houve grande interesse do público no filme durante o período analisado. O filme também não conquistou nenhuma vitória no *Oscar 2017*. Apesar disso, mais da metade dos *tweets* sobre o filme foi classificada com sentimento "positivo". *Hell or High Water* ocupa o penúltimo lugar no *ranking* da Tabela 18, fato que coincide com essa análise.

Um dos filmes com melhor desempenho foi *Hacksaw Ridge*, que obteve o maior número de *tweets* classificados como "positivo" dentre seus *tweets* (Tabela 17). Além do mais, 13% das vitórias dentre o grupo dos nove filmes foram conquistadas por esse filme. Apesar disso, apenas 5,5% da base completa de *tweets* é composta por instâncias relativas a *Hacksaw Ridge*. Como pode-se esperar, o filme ocupa uma posição boa no *ranking* exibido na Tabela 18, o que corresponde com a análise feita sobre essas duas Tabelas.

Os resultados encontrados na Tabela 17 para o filme *Moonlight*, que venceu a categoria Melhor Filme e que ocupa o segundo lugar no *ranking* mostrado na Tabela 18, também têm relação com o que foi encontrado a partir da análise. O longa-metragem teve o terceiro maior índice de *tweets* classificados como "positivo" dentre os seus próprios, ficou entre os cinco com maior quantidade de *tweets* que compõem a base completa e é o quinto filme com o maior número de *tweets* positivos dentre todos os positivos da base.

A partir da análise da Tabela 17, também pode-se perceber que o filme *Arrival* foi o sexto com maior número de *tweets* classificados como "positivo" em relação aos seus próprios, e foi o quarto filme com maior número de *tweets* na base completa e com maior número de *tweets* classificados como "positivo" dentre todos os positivos da base. *Arrival* ocupa o quinto lugar no *ranking* exposto na Tabela 18, uma posição entre as encontradas a partir da análise da Tabela.

Há certa correspondência entre o *ranking* mostrado na Tabela 18 e a opinião dos usuários do *Twitter* (Tabela 17) em relação ao filme *Lion*. O filme foi o sétimo com maior número de *tweets* classificados como "positivos" dentre os seus próprios, mesma posição que ocupou no *ranking* do *Oscar 2017*. *Lion* também é o segundo filme com maior número de *tweets* compondo a base completa e o terceiro com o maior número de *tweets* classificados como "positivo" dentre todos os positivos que compõem a base.

Apesar de ser o terceiro filme com maior quantidade de *tweets* compondo a base completa e de ter mais da metade desses *tweets* classificados como "positivo", o filme *Hidden Figures* obteve o menor número de indicações dentre os filmes avaliados e não venceu em nenhuma categoria. Dessa forma, *Hidden Figures* ocupa a última posição no *ranking* do *Oscar 2017*, apresentado na Tabela 18. Esse é um dos primeiros indícios encontrados de que talvez a opinião da Academia de Ciências e Artes Cinematográficas não represente a opinião do público do *Twitter*.

Fences é o terceiro filme com menor número de *tweets* que compõem a base completa e é um dos dois filmes a não obter pelo menos metade de seus *tweets* classificada como "positivo", de acordo com os dados da Tabela 17. Dessa forma, também há um leve desarcordo entre a opinião dos usuários do *Twitter* e a dos membros da Academia, já que o filme assumiu uma posição maior no *ranking* da segunda coluna da Tabela 18.

Esse desacordo também ocorre em relação ao filme *Manchester by the Sea*, que ocupa o terceiro lugar no *ranking* do *Oscar 2017* (Tabela 18), mas que é o segundo filme com menos *tweets* compondo a base completa e ocupa a oitava posição no *ranking* de acordo com a quantidade *tweets* positivos dentre todos os positivos da base.

4.6.2 Análise da correlação de *Spearman* entre os dados

Com o objetivo de estabelecer a correlação entre os *rankings* do *Oscar 2017* (Tabela 18) e o sentimento dos *tweets*, representados por três *rankings* diferentes (Tabela 17), a correlação de postos de *Spearman*[4] foi utilizada.

Dessa forma, calculou-se o coeficiente de correlação de postos de *Spearman* entre cada um dos *rankings* expostos na Tabela 17 e cada um dos *rankings* mostrados na Tabela 18. Os resultados obtidos são expostos na Tabela 19, com a intenção de descobrir se há alguma associação significativa entre os *rankings*.

	Ranking do Oscar 2017	Indicações	Vitórias
Base completa	0,15	0,43	0,12
Positivos (dentre os próprios)	-0,23	-0,35	-0,12
Positivos (dentre todos os positivos)	0,11	0,36	0,12

Tabela 19 – Valores calculados utilizando a correlação de *Spearman*, comparando os indicadores obtidos a partir do sentimento expresso no *tweets* e do resultado do *Oscar 2017*.

Como é possível perceber, estatisticamente não há nenhuma associação significativa entre os *rankings*. Os maiores valores do coeficiente foram encontrados entre o *ranking* de indicações e os *rankings* de base completa e positivos (dentre todos os positivos), ou seja, a indicação dos filmes movimenta a discussão, mas não necessariamente o que o público pensa sobre os filmes reflete no gosto da Academia.

Dessa forma, é possível perceber que, apesar de nenhum dos coeficientes encontrados representar correlação significativa entre os *rankings*, ainda é bastante provável de

[4] http://www.socscistatistics.com/Default.aspx

se predizer qual filme seria o grande vencedor da premiação – no caso, *La La Land* – e quais filmes estariam entre os menos prestigiados – *Fences*, *Hell or High Water* e *Lion*, por exemplo. Em outras palavras, a partir desta metodologia, é possível predizer as extremidades do *ranking*, apesar de a correlação entre os *rankings* ser baixa. Algumas outras observações também puderam ser realizadas sobre os dados e algumas correspondências encontradas a partir da análise intuitiva.

4.7 Considerações finais

Neste capítulo, uma visão geral do *Oscar 2017* foi apresentada, incluindo informações sobre as indicações e vitórias que cada um dos filmes acumulou e como esses valores formaram o *ranking* do *Oscar 2017* que foi utilizado como referência neste trabalho. Também foi realizada uma análise sobre a quantidade de *tweets* que foram coletados e como essa quantidade pode ter sido afetada devido às diferentes datas de estreia dos filmes. Observações em relação à quantidade de *tweets* coletados por semana e os prováveis motivos, como euforia devido à divulgação dos filmes indicados, também foram feitas.

Após avaliação dos três classificadores por meio da construção de matrizes de confusão e cálculo da acurácia, foi possível escolher o *Naive Bayes* multinomial como o melhor algoritmo de classificação para ser utilizado neste trabalho. Em seguida, o classificador foi aplicado sobre a base completa de *tweets* e foi possível predizer os sentimentos dos usuários do *Twitter* em relação a cada um dos filmes.

A partir dos resultados obtidos, foi possível concluir que a boa parte dos usuários da rede prefere utilizá-la opinar sobre filmes que gostou, já que a quantidade de *tweets* classificados como "negativo" foi inferior ao número de *tweets* classificados como "positivo" para todos os filmes analisados. Além disso, foram encontradas algumas correspondências entre o *ranking* do *Oscar 2017* e os resultados obtidos com o classificador – mesmo que, matematicamente, nenhum dos *rankings* construídos a partir dos resultados teve uma associação significativa com o *ranking* do *Oscar 2017*, o que indica que a opinião da Academia de Artes e Ciências Cinematográficas pode não ter relação direta com o gosto do público em geral.

5 Conclusão

Neste trabalho, foi apresentado um breve histórico sobre as redes sociais e sua importância em diferentes aspectos, como *marketing*, indústria de entretenimento e manifestações sociais. O *Twitter* recebeu foco por ser uma rede social interessante de ser explorada nas áreas de Mineração de Dados e Análise de Sentimentos.

Dessa forma, o objetivo deste trabalho foi realizar a análise dos sentimentos dos *tweets* relacionados aos filmes indicados à categoria de Melhor Filme do *Oscar 2017*. Para isso, foram realizadas as etapas da tarefa de Análise de Sentimentos voltada para o *Twitter*: coleta de *tweets*, pré-processamento dos dados, construção de uma base de dados rotulada, classificação dos textos e validação dos resultados. Três abordagens para classificação de textos foram estudadas e aplicadas na base de dados rotulada: aprendizado supervisionado – utilizando o algoritmo *Naive Bayes* –, aprendizado por supervisão à distância – utilizando a ferramenta *Sentiment140* – e função de polaridade – utilizando a biblioteca *TextBlob*.

Após avaliar o nível de acurácia de cada classificador em relação à classificação da base de dados rotulada, o *Naive Bayes* multinomial foi escolhido para classificar a base completa. Concluiu-se se esse classificador pode ser bastante indicado para tarefas semelhantes, uma vez que obteve um nível de acurácia de quase 75%. A partir dos resultados, foi possível concluir que esse método é interessante para realizar observações sobre os filmes indicados ao *Oscar*, sendo bastante provável de se predizer qual filme seria o grande vencedor da premiação – no caso, *La La Land* – e quais filmes estariam entre os menos prestigiados – *Fences*, *Hell or High Water* e *Lion*, por exemplo.

Também foi possível concluir que boa parte dos usuários do *Twitter* prefere usar a rede para publicar comentários positivos sobre filmes, ao invés de falar mal sobre outros que não gostaram. Os usuários também ficaram bastante animados e ativos na rede social durante a semana em que foram divulgados os indicados ao *Oscar 2017*, mas essa animação diminuiu com o tempo e aumentou moderadamente nas semanas mais próximas do dia da premiação.

Entretanto, estatisticamente não foi encontrada uma associação significativa entre o *ranking* do *Oscar 2017* e os outros *rankings* criados a partir dos resultados obtidos pelo classificador. Isso significa que os dados obtidos precisariam ser interpretados mais profundamente com o objetivo de obter conclusões mais satisfatórias, ao invés de apenas utilizar interpretações matemáticas. Uma outra explicação seria que os filmes que agradam ao público nem sempre são aqueles escolhidos como os melhores pela Academia de Artes e Ciências Cinematográficas – um filme que exemplifica isso é *Hidden Figures*, que foi

muito comentado positivamente pelo público do *Twitter*, mas que recebeu apenas três indicações e não conquistou nenhuma vitória.

5.1 Principais contribuições

- Construção de base de dados sobre filmes indicados à categoria de Melhor Filme do *Oscar 2017*, composta por 889.840 *tweets* pré-processados e separados por filme.
- Construção de base de dados sobre filmes indicados à categoria de Melhor Filme do *Oscar 2017*, composta por 3235 *tweets* rotulados com os sentimentos "positivo", "negativo" ou "neutro". Essa base pode ser usada como conjunto de treinamento para algoritmos de aprendizagem supervisionada.
- Criação de medida para construir um *ranking* do *Oscar* baseada no número de indicações e vitórias conquistadas por cada filme analisado.
- Desenvolvimento de uma ferramenta em *Java* capaz de realizar diversas etapas de pré-processamento de uma base de dados e que pode ser utilizada em outras aplicações.
- Indicação do classificador *Naive Bayes* como promissor para a classificação do sentimento de *tweets*, podendo ser melhor investigado em outros trabalhos.
- Análise da correlação entre o sentimento expresso na rede social *Twitter* e os filmes indicados à categoria de Melhor Filme do *Oscar 2017* sob diferentes aspectos.

5.2 Trabalhos futuros

Neste trabalho, o título de cada filme indicado à categoria de Melhor Filme do *Oscar 2017* definiu a consulta no momento da coleta de *tweets*. Uma análise mais profunda pode ser realizada, considerando as outras categorias da premiação, possibilitando predizer o sentimento dos usuários do *Twitter* em relação aos melhores atores e atrizes, melhores diretores, melhores músicas, entre outros.

Além disso, uma predição mais minuciosa pode ser realizada obtendo-se uma base mais rica, composta não só por *tweets*, mas também por comentários publicados nas redes sociais especializadas em filmes – como o *IMDb* –, avaliações realizadas por críticos ou usuários de outras redes sociais, entre outros.

Uma outra forma de se comparar os resultados do classificador com o resultado do *Oscar* seria analisando apenas certas datas, por exemplo, considerando somente os

tweets publicados na semana em que são divulgados os indicados ou apenas aqueles publicados na semana que antece a premiação. Uma linha do tempo poderia ser construída, demonstrando como os usuários estão se expressando durante cada época escolhida.

Outra abordagem seria considerar apenas as *hashtags* contidas em cada *tweet* e verificar se as mesmas contêm certa emoção que reflete no respectivo *tweet*.

Como foi citado neste trabalho, o perfil do usuário pode influenciar na qualidade da informação. Em um trabalho futuro, seria interessante analisar esses diferentes perfis presentes nas redes sociais e considerar apenas os *tweets* publicados pelos usuários com perfil desejado.

Ademais, a metodologia aplicada neste estudo também pode ser aplicada para futuras premiações ou outras situações em que deseja-se obter um panorama sobre a opinião de usuários do *Twitter* em relação a certo (s) tópico (s). Uma outra forma de abordagem seria desenvolver este estudo antes do resultado do *Oscar* (ou de outra premiação), isto é, sem obter um *ranking* do *Oscar*, realizar conclusões sobre o resultado obtido a partir do classificador escolhido e observar se a conclusão corresponde ao resultado da premiação.

Também seria interessante realizar um estudo utilizando uma base de dados composta por *tweets* publicados após o *Oscar* e analisar o sentimento dos usuários da rede social em relação aos filmes vencedores escolhidos pela Academia.

A Processo de *ground truth* para construção da base rotulada

Após a obtenção da base de dados para cada filme, separada por semana, foi possível realizar a construção da base de *tweets* rotulados. Para isso, foram realizadas as seguintes etapas para a base de cada um dos filmes.

1. Seleção aleatória de *tweets* dentre os publicados em cada uma das cinco semanas analisadas (Figura 28)

Figura 28 – Base de *tweets* do filme *Moonlight* separada em cinco arquivos de acordo com a semana em que os *tweets* foram publicados.

2. Cada *tweet* selecionado aleatoriamente deveria satisfazer as seguintes condições para compor a base rotulada.

   a) O *tweet* está escrito em Inglês.

   b) O *tweet* realmente faz referência ao filme em questão.

   c) O *tweet* contém termos relevantes à classificação e não apenas *links*, *stop words*, etc.

   Esta etapa auxiliou na construção das listas de palavras não relacionadas (Figura 29) que foram utilizadas durante o pré-processamento (Seção 3.3), com o objetivo de remover os *tweets* que não faziam referência aos filmes.

3. Satisfeitas as três condições, o *tweet* era rotulado manualmente pelo autor do trabalho, com o sentimento “positivo”, “negativo” ou “neutro” (Figura 30).

4. Os *tweets* que tiveram sucesso na etapa anterior foram movidos para a base rotulada, junto de seus respectivos rótulos.

i was about to say la la land sounds like a demi lovato song but then i realized it actually is
la la land got so many nominations demi lovato must be so happy since that song came out how long ago? props to her"

Figura 29 – A análise dos *tweets* aleatórios tornou possível perceber, por exemplo, que *tweets* incluindo o nome da cantora Demi Lovato não fazem referência ao filme *La La Land*, pois são brincadeiras com o título do filme e a música de mesmo nome cantada pela artista.

p	Hidden Figures is soooooo good!!!
p	So I was watching hidden figures great movie btw
p	Just watched #HiddenFigures at the cinema. Such a beautiful empowering film about black women who made history
p	hidden figures is sooooo fcking good. i love Taraji
p	I want to see Hidden Figures it looks really good.
p	Just watched Hidden Figures " and wow"

Figura 30 – Alguns *tweets* publicados sobre o filme *Hidden Figures* classificados como positivos.

Referências

ALMEIDA, R. J. d. A. Estudo da ocorrência de cyberbullying contra professores na rede social Twitter por meio de um algoritmo de classificação Bayesiano. *Texto Livre: Linguagem e Tecnologia*, v. 5, n. 1, p. 1–7, 2012. Citado 4 vezes nas páginas 13, 23, 24 e 31.

ARAÚJO, M. et al. Métodos para Análise de Sentimentos no Twitter. *WebMedia '13 Proceedings of the 19th Brazilian symposium on Multimedia and the web*, p. 97–104, 2013. Citado 2 vezes nas páginas 22 e 23.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca*. 2. ed. Porto Alegre: Bookman Editora, 2013. 590 p. Citado 3 vezes nas páginas 22, 24 e 25.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Explorando Redes Sociais Online: Da Coleta e Análise de Grandes Bases de Dados às Aplicações. In: *Livro Texto de Minicursos - SBRC 2011*. Campo Grande: [s.n.], 2011. cap. 2, p. 63–101. Citado 4 vezes nas páginas 13, 16, 20 e 30.

BOTHOS, E.; APOSTOLOU, D.; MENTZAS, G. Using Social Media to Predict Future Events with Agent-Based Markets. *IEEE Intelligent Systems*, v. 25, n. 6, p. 50–58, 2010. Citado 2 vezes nas páginas 11 e 20.

BRAZIL, M. *MPA Brazil*. 2017. <http://www.mpaamericalatina.org/en/>. [Acesso em 22 Abr. 2017]. Citado na página 12.

CERVI, C. R. Um Estudo sobre Mineração de Dados em Redes Sociais. 2008. Citado 2 vezes nas páginas 13 e 30.

CETINSOY, A. *Predicting the 2017 Oscar Winners with BigML*. 2017. <https://dzone.com/articles/predicting-the-2017-oscar-winners>. [Acesso em 19 Nov. 2017]. Citado na página 31.

COSTA, C. *Brasileiros "descobrem"mobilização em redes sociais durante protestos*. 2013. <http://www.bbc.com/portuguese/noticias/2013/07/130628{_}protestos{_}redes{_}personagens{_}cc>. [Acesso em 21 Mai. 2017]. Citado 2 vezes nas páginas 11 e 18.

De Smedt, T.; DAELEMANS, W. "Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 3568–3572, 2012. Citado 3 vezes nas páginas 22, 26 e 27.

DEEPDIVE. *Distant Supervision*. 2017. <http://deepdive.stanford.edu/distant_supervision>. [Acesso em 01 Nov. 2017]. Citado na página 25.

DONNELLY, J. *Watch the Oscar nominations 2017 announcement live!* 2017. <http://oscar.go.com/news/nominations/watch-oscar-nominations-2017-announcement-live>. [Acesso em 01 Mar. 2017]. Citado 2 vezes nas páginas 34 e 48.

FACEBOOK. *Company Info*. 2017. <https://newsroom.fb.com/company-info/>. [Acesso em 07 Jun. 2017]. Citado na página 16.

FELIX, N. *Análise de sentimentos em textos curtos provenientes de redes sociais*. 138 p. Tese (Doutorado) — Universidade de São Paulo - São Carlos, 2016. Citado 7 vezes nas páginas 18, 20, 21, 23, 24, 35 e 36.

FILHO, J. A. C. Mineração De Textos: Análise de Sentimento Utilizando Tweets Referentes à Copa Do Mundo 2014. 2014. Citado 11 vezes nas páginas 11, 12, 13, 17, 18, 23, 24, 28, 29, 31 e 37.

G1. *Cidades têm domingo de protestos contra Dilma e contra Temer*. 2016. <http://g1.globo.com/politica/noticia/2016/07/cidades-tem-domingo-de-protestos-contra-dilma-e-contra-temer.html>. [Acesso em 21 Mai. 2017]. Citado 2 vezes nas páginas 11 e 18.

GIGLIOTTI, W. Em Busca de Más Notícias. 2012. Citado 2 vezes nas páginas 27 e 41.

GO, A.; BHAYANI, R.; HUANG, L. Twitter Sentiment Classification using Distant Supervision. *Processing*, v. 150, n. 12, p. 1–6, 2009. Citado 3 vezes nas páginas 23, 26 e 35.

INSTAGRAM. *Our Story*. 2017. <https://instagram-press.com/our-story/>. [Acesso em 07 Jun. 2017]. Citado na página 17.

KRAUSS, J.; NANN, S.; SIMON, D. Predicting movie success and academy awards through sentiment and social network analysis. In: *16th European Conference on Information Systems*. [S.l.: s.n.], 2008. p. 12. ISBN 978-0-9553159-2-3. Citado 3 vezes nas páginas 13, 31 e 32.

LIU, B. *Sentiment Analysis and Opinion Mining*. Chicago: Morgan & Claypool Publishers, 2012. 168 p. Citado 2 vezes nas páginas 19 e 20.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ETMTNLP '02), p. 63–70. Citado na página 24.

LORIA, S. et al. *TextBlob: Simplified Text Processing*. 2017. <http://textblob.readthedocs.io/en/dev/index.html>. [Acesso em 28 Ago. 2017]. Citado 2 vezes nas páginas 26 e 27.

MANDEL, B. et al. A Demographic Analysis of Online Sentiment during Hurricane Irene. In: *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*. Montréal: [s.n.], 2012. p. 27–36. ISBN 9781937284206. Citado 5 vezes nas páginas 6, 13, 18, 29 e 30.

MCHATTON, K. *Infographic: The film industry in numbers*. 2015. <https://www.icas.com/ca-today-news/infographic-the-film-industry-in-numbers>. [Acesso em 22 Abr. 2017]. Citado na página 12.

OLIVEIRA, F. W. C. de. Análise de sentimentos de comentários em português utilizando SentiWordNet. 2013. Citado 5 vezes nas páginas 11, 20, 21, 23 e 24.

PAK, A.; PAROUBEK, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation*. [S.l.: s.n.], 2010. p. 1320–1326. ISBN 2951740867. Citado na página 31.

PATEL, N. *Redes Sociais: O Guia Completo para Definir Suas Estratégias de Marketing*. 2015. <http://neilpatel.com/br/redes-sociais-o-guia-completo-para-definir-suas-estrategias-de-marketing/>. [Acesso em 08 Jun. 2017]. Citado na página 17.

PAYNTER, G. et al. *Attribute-Relation File Format (ARFF)*. 2002. <https://www.cs.waikato.ac.nz/ml/weka/arff.html>. [Acesso em 02 Out. 2017]. Citado na página 40.

REIS, J.; GONÇALVES, P.; ARAÚJO, M. Uma Abordagem Multilıngue para Análise de Sentimentos. *Each.Usp.Br*, p. 12, 2012. Citado na página 22.

RIBEIRO, F. N. et al. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, v. 5, n. 1, 2016. ISSN 21931127. Citado na página 24.

RIBEIRO, L. B. Análise de sentimento em comentários sobre aplicativos para dispositivos móveis: Estudo do impacto do pré-processamento. 2015. Citado 3 vezes nas páginas 23, 24 e 25.

ROSSI, M. *Protestos contra Dilma voltam embalados por escalada da crise política*. 2016. <http://brasil.elpais.com/brasil/2016/03/09/politica/1457553690{_}568304.html>. [Acesso em 21 Mai. 2017]. Citado 2 vezes nas páginas 11 e 18.

SCHMITT, V. F. Uma Análise Comparativa De Técnicas De Aprendizagem De Máquina Para Prever a Popularidade De Postagens No Facebook. 2013. Citado 3 vezes nas páginas 24, 25 e 28.

SEMIOCAST. *Arabic highest growth on Twitter - English expression stabilizes below 40 percent*. 2011. <https://semiocast.com/publications/2011_11_24_Arabic_highest_growth_on_Twitter>. [Acesso em 28 Out. 2017]. Citado na página 35.

SMITH, C. *135 Amazing Snapchat Statistics and Facts (May 2017)*. 2017. <http://expandedramblings.com/index.php/snapchat-statistics/>. [Acesso em 07 Jun. 2017]. Citado na página 17.

SMITH, T. C.; FRANK, E. *Statistical Genomics: Methods and Protocols*. New York, NY: Springer, 2016. 353–378 p. Disponível em: <http://dx.doi.org/10.1007/978-1-4939-3578-9_17>. Citado 2 vezes nas páginas 24 e 40.

SOUZA, G. L. S. de. Tweetmining: Análise de Opinião Contida em Textos Extraídos do Twitter. 2012. Citado 2 vezes nas páginas 23 e 24.

STATS, I. L. *Twitter Usage Statistics*. 2017. <http://www.internetlivestats.com/twitter-statistics/>. [Acesso em 20 Abr. 2017]. Citado 2 vezes nas páginas 12 e 18.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. Citado na página 44.

TAVARES, V. B. A. *O Papel das Redes Sociais na Primavera Árabe de 2011: implicações para a ordem internacional*. 2012. <https://www.mundorama.net/?p=10624>. [Acesso em 22 Abr. 2017]. Citado na página 11.

TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, v. 8, p. 53–65, 2011. Citado 5 vezes nas páginas 13, 16, 18, 31 e 32.

TUMASJAN, A. et al. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. [S.l.: s.n.], 2010. p. 178–185. ISBN 0894439310386. Citado na página 12.

TWITTER. *Company*. 2017. <https://about.twitter.com/company>. [Acesso em 20 Abr. 2017]. Citado 3 vezes nas páginas 11, 17 e 33.

TWITTER. *Twitter milestones: a selection of memorable moments*. 2017. <https://about.twitter.com/company/press/milestones>. [Acesso em 07 Jun. 2017]. Citado na página 17.

WONG, V. *How Oscar Nominations Affect the Box Office*. 2013. <https://www.bloomberg.com/news/articles/2013-01-10/how-oscar-nominations-affect-the-box-office>. [Acesso em 22 Abr. 2017]. Citado na página 12.