UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ALEX DE OLIVEIRA ALEXANDRINO

ANÁLISE DE REDES SOCIAIS APLICADA A TWEETS SOBRE SÉRIES DE TV

QUIXADÁ
2016

ALEX DE OLIVEIRA ALEXANDRINO

ANÁLISE DE REDES SOCIAIS APLICADA A TWEETS SOBRE SÉRIES DE TV

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal do Ceará como requisito parcial para obtenção do Título de Bacharel em Sistemas de Informação.

Orientadora Profª. Paulyne Matthwes Jucá

QUIXADÁ
2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

A371a Alexandrino, Alex de Oliveira

Análise de redes sociais aplicada a tweets sobre séries de tv/ Alex de Oliveira Alexandrino. – 2016.
48 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Bacharelado em Sistemas de Informação, Quixadá, 2016.
Orientação: Profa. Dra. Paulyne Matthews Jucá
Área de concentração: Computação

1. Redes sociais - Análise 2. Televisão - Seriados 3. Twitter (Rede social on line) 4. Mineração de dados (Computação) I. Título.

CDD 004.6

ALEX DE OLIVEIRA ALEXANDRINO

ANÁLISE DE REDES SOCIAIS APLICADA A TWEETS SOBRE SÉRIES DE TV

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel. Área de concentração: computação.

Aprovado em: _____ / fevereiro / 2016.

BANCA EXAMINADORA

_____
Profª. Dr. Paulyne Matthews Jucá (Orientadora)
Universidade Federal do Ceará-UFC

_____
Prof. Dr. Arthur de Castro Callado
Universidade Federal do Ceará-UFC

_____
Prof. MSc. Régis Pires Magalhães
Universidade Federal do Ceará-UFC

A minha família, por todo o apoio e ajuda que me deram durante toda essa jornada da graduação, e acima de tudo por todos os ensinamentos que me passaram na vida, não seria quem eu sou hoje se não fossem eles.

AGRADECIMENTOS

Agradeço aos meus pais, Pedro Alexandrino e Marileide Oliveira, que mesmo nos momentos mais complicados nunca deixaram de me apoiar, e que sempre me ensinaram a ser um homem honesto e de caráter.

Agradeço ao meu irmão, Alexsandro Oliveira, pela amizade e companheirismo que sempre tivemos.

Agradeço a minha namorada e grande amiga Samara Lou, por todo o apoio e ajuda que me deu nesses últimos meses.

Agradeço a todos meus amigos que trouxe de Jucás, Katyeudo, Apolônio, Warniery, Yago, João Victor, Samu, Rodrigo, os quais acompanharam toda a jornada anterior a minha ida para Quixadá e toda a jornada construída em Quixadá.

Agradeço aos meus colegas de apartamento e de sala que aprendi a chamar de amigos, Araújo, Júnior, Sávio, Wanrly, Adeilson, William, Guilherme, Douglas, Matheus, Anderson, Emanuel, Zé Roberto e Danrley. Todos fizeram parte desses quatro anos da graduação.

Agradeço a minha orientadora Paulyne Matthews, que durante todos os meus quatro anos de graduação esteve presente nos meus projetos dentro da universidade, repassando grandes ensinamentos, os quais sempre levarei tanto na vida profissional quanto na vida pessoal.

Por fim, um agradecimento especial, ao professor que me mostrou as primeiras linhas de código, escritas em portugol a lápis em um papel. Agradeço ao Arley Rodrigues, que infelizmente não está mais entre nós, mas tenho certeza que de onde estiver, e sei que está em um lugar bom pela grande pessoa que ele foi, está orgulhoso de toda a jornada feita e tudo que foi construído por mim nela.

"A simplicidade é o último grau de sofisticação." (Leonardo da Vinci)

RESUMO

O grande uso das redes sociais e as diversas formas como elas podem ser acessadas modificaram a forma como pessoas e organizações interagem. Um vasto volume de informações é gerado constantemente. Muitas organizações têm investido na análise desses dados a fim de obter informações sobre seus clientes. A análise de redes sociais e a análise de sentimentos, juntamente com a mineração de textos, fornecem técnicas que podem estruturar e facilitar o entendimento das informações. Dentre as redes sociais mais utilizadas, está o Twitter, *microblog* que permite que pessoas publiquem mensagens, chamadas de *tweets*, expressando suas opiniões, o que a torna uma das principais fontes para a análise de opiniões. Nesse contexto, este trabalho visa coletar *tweets* sobre séries de TV e aplicar um conjunto de técnicas de análise de redes sociais, análise de sentimentos e mineração de textos. As métricas mais utilizadas na literatura para a análise de redes sociais foram escolhidas e aplicadas, possibilitando a análise da quantidade e força da relação entre as séries de TV, assim como da popularidade de cada série. Para a classificação de sentimentos, as mensagens foram categorizadas com um sentimento positivo, negativo ou neutro.

Palavras-chave: Análise de redes sociais. Análise de sentimentos. Twitter.

ABSTRACT

The widespread use of social networks and the various ways they can be accessed changed the way people and organizations interact with each other, generating a vast amount of information. Many organizations have invested in the analysis of such data in order to obtain information about their clients. The social network analysis and sentiment analysis, along with text mining, provide techniques that can structure and facilitate the understanding of information. Twitter is among the most popular social networks, it is a microblog that allows people publish messages, called tweets, expressing their opinions, making it one of the main sources for the analysis of opinions. In this context, this work aims to collect tweets about TV shows and apply a set of techniques of social networks analysis, sentiment analysis and text mining. The metrics most commonly used in the literature for the analysis of social networks were selected and applied, enabling the analysis of the quantity and strength of the relationship between the TV shows, as well as the popularity of each TV show. For the classification of feelings, messages were categorized with a positive, negative or neutral feeling.

Keywords: Social networking analysis. Sentiment analysis. Twitter.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O grande número de redes sociais, o seu crescente uso e o constante avanço tecnológico modificaram as formas de relação entre pessoas e organizações. O que antes acontecia na tarde de uma segunda-feira e era publicado no jornal impresso no dia seguinte, ou noticiado à noite na TV, hoje pode ser publicado em tempo real nos grandes portais de notícias e compartilhado por várias pessoas nas redes sociais.

O termo rede social não é algo novo e também não surgiu no âmbito tecnológico, mas foi somente quando passaram a ser representadas em software que o tema ganhou uma nova perspectiva (MEIRA et al., 2011).

Um vasto volume de informações é gerado a cada instante, atraindo assim a atenção de muitas organizações em torno dos benefícios da análise desses dados. Ao analisar essas informações, as organizações adquirem um conhecimento mais aprofundado das opiniões dos seus clientes sobre seus serviços e produtos (GOMES, 2013), além de criar um poderoso canal de comunicação direta e divulgação da sua marca.

Nesse contexto, a análise de redes sociais (ARS - *network social analysis*) e a análise de sentimentos (AS - *sentiment analysis*), juntamente com a mineração de textos (*text mining*), apresentam algumas técnicas que podem estruturar e facilitar o entendimento de toda essa informação, permitindo assim que as organizações saibam o que as pessoas estão comentando sobre elas em seus perfis nas redes sociais. A partir dessa análise, as organizações podem tirar proveito para elaborar planos de marketing, comunicação com o cliente, melhoria de serviços, dentre outros (CARVALHO FILHO, 2014).

Dentre as redes sociais mais utilizadas, está o Twitter, um *microblog* que permite que pessoas publiquem mensagens de tamanho limitado em 140 caracteres, expressando sua opinião sobre algo ou disponibilizando alguma informação para os seus contatos. Os *tweets* publicados na rede possuem opiniões, informações pessoais ou informações sobre eventos em geral (NAAMAN; BOASE, 2010). Por esse motivo, o Twitter é visto como uma das principais fontes para análise de sentimentos e de opiniões sobre eventos e acontecimentos (LI; LI, 2011).

Séries de TV são um tipo de programa de televisão que possui por temporada um número pré-determinado de episódios. A maioria das emissoras de televisão opta por lançar episódios semanalmente. Algumas séries são lançadas por serviços de streaming, como o

Netflix[1], sendo a temporada lançada toda de uma única vez na maioria das vezes. No atual momento, existe uma grande variedade de séries de TV classificadas em várias categorias. Algumas dessas categorias são drama, ação, suspense e comédia. As redes sociais estão sendo muito usadas para as pessoas compartilharem alguma opinião sobre as séries de TV. É possível saber através das mensagens publicadas nas redes sociais opiniões, popularidade, entre outras informações sobre as séries de TV.

Alguns trabalhos já foram desenvolvidos visando a análise de *tweets.* O número de trabalhos em português ainda é pequeno. Entre alguns dos trabalhos em português estão os de Recuero (2014), Carvalho Filho (2014) e Viana (2014), todos os quais analisaram *tweets* sobre algum tema específico. O primeiro usou técnicas de análise de redes sociais e os dois últimos utilizaram técnicas de análise de sentimentos.

O presente trabalho tem como objetivo avaliar *tweets* referentes a um conjunto de séries de TV aplicando um conjunto de técnicas de análise de redes sociais e de análise de sentimentos.

O primeiro passo do trabalho consistiu na escolha de quais séries seriam analisadas. Em paralelo foram escolhidas as *hashtags* que iriam ser buscadas e foi construído o *crawler* que realizou a coleta. Um *crawler* nada mais é que um algoritmo que realiza buscas na *WEB* por conteúdos relevantes à sua função de forma metódica e automatizada (ARANHA, 2007). Os *tweets* foram coletados basicamente em três datas distintas para quase todas as séries, à exceção de algumas séries lançadas em serviços de *streaming*, como a Netflix, já que as temporadas geralmente são lançadas por completo em único dia. A coleta foi restrita a *tweets* em português brasileiro.

Logo após a base da coleta estar construída, foi realizado um pré-processamento nas mensagens para a retirada de conteúdos julgados irrelevantes para a análise desejada. Em seguida foi construído o modelo de classificação.

Foram criadas nuvens de palavras para um conjunto de séries escolhidas de acordo com a quantidade de *tweets* coletados, com a intenção de demonstrar qual a frequência com que as *hashtags* aparecem nos textos. Por fim, os dados foram avaliados seguindo regras de um conjunto de técnicas de análise de redes sociais e de análise de sentimentos. As métricas de ARS de cálculo do grau do nó, densidade, popularidade, centralidade e grau de proximidade foram as escolhidas para serem usadas no presente trabalho. Foi identificado como as séries se relacionam e qual o papel de cada uma dentro da rede social criada. Também foram

[1] www.netflix.com

identificadas quais as séries possuem maior número de comentários dentro do conjunto de *tweets* coletados. Em relação à aplicação da análise de sentimentos, foi verificada a polaridade (opiniões positivas, negativas ou neutras) expressa nos *tweets*. Por fim, serão avaliados e demonstrados os resultados obtidos.

Este trabalho está organizado conforme segue: o capítulo 2 apresenta um resumo sobre os conceitos técnicos e teóricos necessários para a realização do trabalho. O capítulo 3 contém três trabalhos relacionados. Serão discutidas na seção as propostas dos trabalhos e como eles se assemelham e se diferenciam da proposta aqui apresentada. O capítulo 4 apresenta o objetivo geral e os objetivos específicos do trabalho. O capítulo 5 mostra a descrição do procedimento metodológico que foi realizado no trabalho, contendo todos os passos necessários para atingir os objetivos propostos no trabalho. O capítulo 6 mostra os resultados do trabalho. O capítulo 7 apresenta os trabalhos futuros que serão realizados. Por fim, a Seção 8 apresenta a conclusão do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Nessa seção, serão abordados os principais conceitos relacionados a este trabalho e qual a contribuição de cada conceito para o seu desenvolvimento.

2.1 Redes Sociais

Uma rede social pode ser definida como uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, podendo ser redes reais ou virtuais. Meira et al. (2011) definem redes sociais virtuais como representações computacionais de redes existentes ou criadas a partir da relação de pessoas que, muitas vezes, nunca se conheceram pessoalmente. Dada essa definição, todas as menções ao termo rede social neste trabalho estarão se referindo a rede social virtual.

Golbeck (2005) define alguns critérios fundamentais para que aplicações de redes sociais baseadas na WEB (*Web Based Social Networks* - WBSN) sigam:

- Deve ser completamente acessível via *WEB* – aplicações que precisam realizar um *download* de algum *software* para ser acessível não são consideradas redes sociais. Assim, necessitam apenas de um *browser* para realizar o acesso;
- Os *status* de relacionamento entres as pessoas necessitam estar descritos - Pessoa A é amigo de Pessoa B e Pessoa B é irmão de Pessoa C, por exemplo;
- É necessário um suporte claro e integrado da aplicação para que os usuários possam ser capazes de criar estas conexões e os relacionamentos devem ser visíveis e navegáveis, ou seja, um usuário deverá ser capaz de navegar na lista de amigos do seu amigo.

As redes sociais evoluíram consideravelmente nos últimos anos, principalmente no que diz respeito ao aumento do seu escopo. Há um grande número de redes sociais, com os propósitos mais diversos, envolvendo jogos, músicas, notícias e fotos.

Foi utilizada, no presente trabalho, a rede social Twitter. Foram coletados *tweets* sobre séries de TV, e posteriormente analisados seguindo técnicas de análise de redes sociais e de análise de sentimentos.

2.2 Análise de redes sociais

Meira et al. (2011) define análise de redes sociais (ARS - *Social Network Analysis*) como uma área que tem como finalidade coletar e analisar padrões nos relacionamentos e fluxos de informações existentes entre os nós (pessoas, grupos, organizações, entre outros) de uma

rede social, tornando possível conseguir informações visuais e matemáticas dos relacionamentos dos nós. Os nós representam entidades que possuem informação ou processo de conhecimento. A ARS pode ser aplicada a várias áreas de conhecimento como em pesquisas sobre economia, recomendação de produtos, gerenciamento organizacional e análise de satisfação do cliente. O resultado da ARS pode ajudar a identificar qual papel cada indivíduo tem na rede social em questão.

Quando a ARS é usada em redes que representam empresas, ela também pode ser nomeada como análise de rede organizacional (ARO - *Organization Network Analysis*).

A seguir serão apresentadas as principais métricas utilizadas na ARS.

2.2.1 Métricas

Um nó representa um elemento de uma rede social, em que sua conexão é representada por uma linha que liga esses dois nós. Assim, uma rede social pode ser representada visualmente por um grafo que representa todas as conexões de um nó com os demais nós, em que essas conexões podem ser direcionais (possuir um sentido) ou não direcionais. As conexões, ou relacionamentos, também podem ser representadas com pesos, indicando a importância de uma conexão. Quanto maior for o valor do peso, mais forte é a conexão entre os nós (MEIRA et al., 2011).

Alguns indicadores são utilizados para realizar uma análise sobre as características de uma rede social. Dentre as principais métricas, temos:

- Grau do nó: representa o número de conexões que um determinado nó possui. Um grafo direcionado tem dois graus: o *in-degree*, que representa o número de conexões que um nó recebe, e o *out-degree*, que representa o número de conexões que saem de um nó. Para um grafo não-direcionado, existe apenas um grau, que é o número de conexões que ele possui (PASSMORE, 2011);
- Grau de proximidade: indica o quão perto um nó está do restante dos nós. É calculada medindo todas as distâncias geodésicas de um nó para se ligar aos restantes. Uma distância geodésica pode ser definida como a menor distância que liga dos pontos (PASSMORE, 2011);
- Grau de centralidade: quantidade de nós aos quais um nó está diretamente relacionado (PASSMORE, 2011);

- Densidade: medida calculada através da proporção de relacionamentos existentes no nó ou na rede em relação à quantidade máxima de possíveis relacionamentos para o nó ou para a rede (MEIRA et al., 2011);
- Popularidade: medida pela quantidade de relacionamentos que um nó possa ter com os demais nós (MEIRA et al., 2011);

Existem outras métricas para realizar ARS, porém as que se tornaram de valor para a construção do presente trabalho foram as citadas, e todas foram usadas na presente proposta.

No grafo que foi gerado, as séries de TV foram os nós, as relações entre as séries foram as arestas e os pesos das arestas foram definidas de acordo com a quantidade de menções que uma relação entre duas séries possuem. Por exemplo, se as séries *Game of Thrones* e *Demolidor* são citadas pelos usuários em uma mesma mensagem 50 vezes, o peso dessa relação será igual a 50. Graficamente, os nós são representados por círculos, as arestas por linhas que ligam dois nós e o peso da aresta é representado pela largura da linha da aresta, ou seja, quanto maior o peso da aresta, mais larga será a linha que lhe representa.

As relações entre duas séries foram definidas no presente trabalho como uma citação delas em um *tweet* de um usuário, por exemplo, se um usuário escreve um *tweet* com o texto "Estou assistindo *Demolidor*, mais tarde irei assistir o novo episódio de *Game of Thrones*", as séries *Game of Thrones* e *Demolidor* possuem uma relação.

2.3 Mineração de textos

"A Mineração de Textos, também conhecida como Descoberta de Conhecimento de Texto refere-se ao processo de extrair padrões interessantes e não triviais ou conhecimento a partir de textos desestruturados. " (TAN, 1999, p.1, tradução livre)

Aranha (2007) propõe cinco grandes etapas para compor um processo de mineração de texto (Figura 1): coleta, pré-processamento, indexação, mineração e análise da informação.

A coleta tem como objetivo compor a base textual de trabalho com informações extraídas do meio externo. Essas informações geralmente são coletadas utilizando um *crawler*. Um *crawler* pode ser definido como um algoritmo que realiza buscas na *WEB* por conteúdos relevantes à sua função de forma metódica e automatizada (ARANHA, 2007). Nessa fase, determina-se qual será o universo de aplicação das técnicas de mineração de textos.

O pré-processamento é aplicado a um conjunto de técnicas para realizar a estruturação dos textos em um formato adequado para serem submetidos pelos algoritmos de análise, ou seja, melhorar a qualidade e organizar os dados já disponíveis. Entre algumas técnicas aplicadas, temos a retirada de *stopwords* e a aplicação de *steaming*. *Stopwords* são

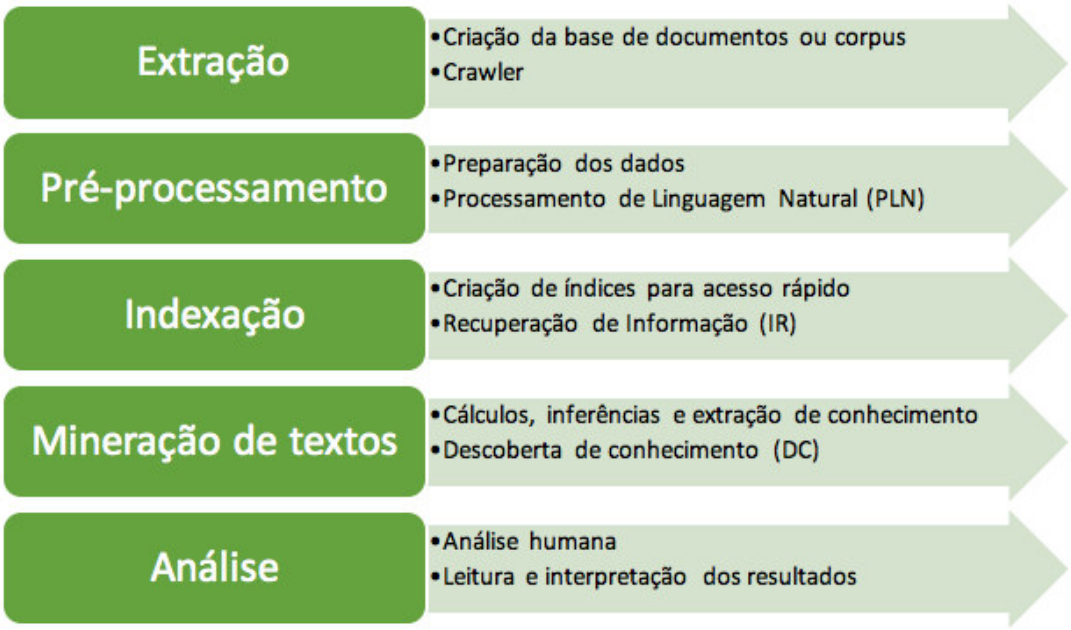
palavras consideradas desnecessárias em uma mensagem. Alguns exemplos são artigos, preposições e pequenas palavras. *Steaming* é a operação da retirada do radical de uma palavra.

A indexação consiste na criação de índices para acesso rápido, ou seja, permite que a procura/recuperação de informações seja realizada de forma mais otimizada.

A mineração é a etapa que envolve a escolha de quais algoritmos serão aplicados aos dados. Essa escolha depende do objetivo de estudo, podendo assim utilizar algoritmos provenientes de diversas áreas, como aprendizado de máquina, banco de dados e redes neurais.

Por fim, a fase de análise da informação consiste da avaliação dos resultados obtidos. Esta serve tanto para validar os algoritmos utilizados, quanto para verificar o conhecimento extraído (GOMES, 2013).

Figura 1 – Etapas do processo de Mineração de Textos



Fonte: adaptada de Aranha (2007)

Foram realizadas no presente trabalho todas as etapas descritas acima, com enfoque maior nas de coleta, pré-processamento, mineração e análise de informações. A fase de indexação não será importante para este trabalho, uma vez que o trabalho não prioriza a melhoria de desempenho que poderia ser obtida ao usar índices.

Foi utilizado na fase de mineração o algoritmo de aprendizagem estatística Naive Bayes para criação do modelo de classificação. O Naive Bayes é um classificador probabilístico baseado na aplicação de Bayes.

Figura 2 – Teorema de Bayes

$$P[H \mid E] = (P[E \mid H]\ P[H]) / P[E]$$

Fonte: elaborada pelo autor.

A Figura 2 ilustra o Teorema de Bayes, onde E representa um evento que ocorreu previamente, e H é um evento que depende de E. É calculada a probabilidade de H ocorrer dado o evento E. O algoritmo deverá contar o número de casos em que H e E ocorrem juntos e dividir pelo número de casos em que E ocorre sozinho.

Algumas métricas são utilizadas para avaliar o modelo de classificação considerando os valores apresentados na Tabela 1, sendo *a, e* e *i* o número de mensagens classificadas corretamente, respectivamente, como positivas (*true positive*), negativas (*true negative*) e neutras (*true neutral*); *d* e *g* o número de mensagens positivas classificadas, respectivamente, como negativas (*false negative*) e neutras (*false neutral*); *b* e *h* o número de mensagens negativas classificadas, respectivamente, como positivas (*false positive*) e neutras (*false neutral); c* e *f* o número de mensagens neutras classificadas, respectivamente, como positivas (*false positive*) e negativas (*false negative*).

Tabela 1: Matriz de confusão

		Observação real		
		Positivo	Negativo	Neutro
Predição esperada	Positivo	*a*	*b*	*c*
	Negativo	*d*	*e*	*f*
	Neutro	*g*	*h*	*i*

Fonte: elaborada pelo autor.

Para realizar a avaliação do modelo, foram consideradas as seguintes métricas:

- Acurácia - calculada para o modelo. Seu cálculo funciona da seguinte forma: $A = (a + e + i) / (a + b + c + d + e + f + g + h + i)$;
- Precisão - calculada para uma classe. Por exemplo, para a classe positiva, seu cálculo é: $P_{pos} = a / (a + b + c)$;
- *Recall* - calculada para uma classe. Por exemplo, para a classe positiva, seu cálculo é: $R_{pos} = a / (a + d + g)$;
- *F-score* – calculada para uma classe. Por exemplo, para a classe positiva, seu cálculo é: $F_{pos} = 2 * (P_{pos} * R_{pos}) / (P_{pos} + R_{pos})$.

O algoritmo *bayesiano* utilizado neste trabalho é uma implementação pertencente ao *Rapidminer*[2]. De acordo com Gomes (2013), na fase de classificação geralmente são utilizados os algoritmos de Bayes, por possuírem facilidade na implementação e com eles se obter bons resultados. Por fim, foram avaliados e demonstrados os resultados obtidos.

[2] https://rapidminer.com/

3 TRABALHOS RELACIONADOS

O presente trabalho baseia-se em informações, métodos, conceitos e técnicas coletadas de livros, artigos e trabalhos científicos coletados das áreas de análise de redes sociais, análise de sentimentos e mineração de textos. Os trabalhos de Recuero (2014), Carvalho Filho (2014) e Viana (2014) são os principais contribuidores para o desenvolvimento deste trabalho.

Assim como nestes três trabalhos citados, o trabalho aqui proposto irá realizar uma coleta de mensagens na rede social Twitter. Entretanto, o trabalho aqui proposto não realizou apenas análise de sentimentos, mas sim aplicou também algumas técnicas de análise de redes sociais. Além disso, o foco será nos *tweets* publicados relacionados a um conjunto de séries de TV, diferente dos temas propostos nos trabalhos citados. Outro ponto que os dois últimos trabalhos apresentam e que este trabalho realizou é a criação de nuvens de palavras. A Tabela 2 apresenta as diferenças e semelhanças entre os trabalhos propostos e o presente trabalho.

Tabela 2 – Diferenças e semelhanças dos trabalhos relacionados com o trabalho proposto

	AS	**ARS**	**Nuvens de palavras**	**Coleta no Twitter**	**Tema**
Recuero (2014)	Não	Sim	Não	Sim	Protestos que ocorreram no Brasil entre junho e agosto de 2013
Carvalho Filho (2014)	Sim	Não	Sim	Sim	Copa do Mundo de 2014
Viana (2014)	Sim	Não	Sim	Sim	Eleições presidenciais no Brasil em 2014
Presente Trabalho	Sim	Sim	Sim	Sim	Séries de TV

Fonte: elaborada pelo autor.

A seguir veremos o que cada um destes trabalhos aborda.

3.1 Contribuições da Análise de Redes Sociais para o estudo das redes sociais na Internet: o caso das hashtags #Tamojuntodilma e #CalaabocaDilma

Recuero (2014) propôs, no seu trabalho, discutir quais as contribuições que a aplicação das técnicas de análise de redes sociais (ARS) fornece para o estudo de redes sociais

na Internet. Para isso, é realizada uma análise de duas *hashtags* relacionadas aos protestos que aconteceram no Brasil em junho de 2013.

Na primeira parte do trabalho, Recuero (2014) mostra quais são as principais métricas utilizadas para a ARS. Logo após é realizada uma coleta de textos na rede social Twitter, buscando pelas duas *hashtgs* que serão analisadas, que são #tamojuntodilma e #calabocadilma. Logo após, foram aplicadas técnicas de ARS nos *tweets.*

O presente estudo demonstrou como a ARS permite identificar padrões nos dados que podem ilustrar os contextos e as dinâmicas dos grupos que se manifestam.

3.2 Mineração de textos: análise de sentimentos utilizando *tweets* referentes à Copa do mundo de 2014

Carvalho Filho (2014) propôs, no seu trabalho, categorizar os *tweets* de acordo com o sentimento expresso acerca da Copa do Mundo de futebol, realizada em 2014. A Copa do Mundo é realizada pela FIFA, sendo a maior competição de futebol realizada no mundo, tendo sua realização no ano de 2014 ocorrida no Brasil. Carvalho Filho (2014) defende a escolha desse cenário baseado na grande cobertura apresentada e no interesse social que a competição desperta na população brasileira.

Sua execução foi iniciada com a coleta dos *tweets*. Foi utilizado um script escrito na linguagem Python, o qual recebia como parâmetro uma lista de *hashtags* e tinha como retorno *tweets* que continham essas *hashtags*. O início do processo aconteceu dia 12 de junho, início da copa, e foi finalizado no dia 13 de julho, final da copa. Logo após a coleta dos dados foi realizado o pré-processamento dos dados, a fim de descartar informações dos textos considerados irrelevantes para o processo de classificação dos textos, como links e nomes de usuário, por exemplo. Logo após, o algoritmo de classificação de textos Naive Bayes foi utilizado para gerar o modelo de classificação. Por fim, foram aplicadas algumas técnicas de análise de sentimentos na base de *tweets*.

Foi possível perceber após a análise que o conteúdo presente nas mensagens coletadas refletia o sentimento e o pensamento da população brasileira durante o evento.

3.3 Mineração de textos: Análise de Sentimento utilizando *tweets* referentes às eleições presidenciais 2014

Viana (2014) propôs no seu trabalho analisar a opinião da população brasileira sobre as eleições para Presidente da República, utilizando os *tweets* publicados na rede social Twitter durante o período das eleições presidenciais (agosto, setembro e início de outubro).

O primeiro passo para a execução do trabalho foi a escolha de um conjunto de candidatos presidenciáveis a serem analisados. Logo após foram escolhidas as *hashtags* que iriam ser analisadas. Em seguida, foi realizada a coleta dos dados, utilizando um algoritmo escrito na linguagem Ruby[3], utilizando a API disponibilizada pelo Twitter. Após a base estar pronta, foi realizado um pré-processamento nos dados a fim de eliminar informações irrelevantes para a classificação. Logo após foi aplicado o algoritmo de classificação Naive Bayes nos textos pré-processados, sendo que após essa etapa, foram aplicadas algumas técnicas de análise de sentimentos.

Foi possível, no fim, avaliar os percentuais de cada tipo de sentimento em períodos distintos para cada candidato avaliado no trabalho.

[3] www.ruby-lang.org/

4 OBJETIVOS

4.1 Objetivo geral

Avaliar *tweets* sobre séries de TV aplicando técnicas de análise de redes sociais e análise de sentimentos.

4.2 Objetivos específicos

- Identificar as técnicas normalmente utilizadas para analisar conteúdo de redes sociais;
- Escolher as técnicas de análises de redes sociais que serão utilizadas;
- Desenvolver um *crawler* para realizar a coleta dos *tweets* da rede social Twitter;
- Coletar e analisar os dados da rede social Twitter sobre séries de TV;
- Criar nuvens de palavras para um conjunto de séries.

5 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, serão descritas todas as fases da execução do trabalho.

5.1 Identificação das séries de TV

O primeiro passo para o desenvolvimento do trabalho foi a escolha de quais séries de TV seriam analisadas. Alguns pontos foram levados em consideração para essa escolha. O primeiro ponto foi o quão comentada uma série é no Twitter, excluindo da coleta séries que possuiam um baixo índice de comentários no Twitter, pois a quantidade de *tweets* não seria relevante para realizar uma análise adequada. Séries com menos de 10 mil *tweets* foram desconsideradas. O segundo ponto foi a época de término da temporada de uma série. Foram incluidas no trabalho séries que possuíam término até outubro de 2015, pois o tempo limite para o término do presente trabalho não permitiu ir mais adiante.

Foram escolhidas quatorze séries, que são: *Demolidor*, *Game of Thrones*, *Orphan Black*, *Orange is the New Black, Hannibal, Sense 8, Narcos, True Detective*, *Under the Dome*, *Penny Dreadful, Suits, The Strain, Between* e *We Hot American Summer*.

5.2 Desenvolvimento do *crawler*

Para o desenvolvimento do *crawler* foi utilizada a API disponibilizada pelo Twitter, já que é através do seu uso que será possível obter a autorização para realizar as buscas e definir como realizar as buscas. Foi usada como linguagem de desenvolvimento o Java. Sua escolha se deve pela robustez da linguagem, facilidade de manutenção de código e o tempo gasto para o desenvolvimento da aplicação ser menor.

A API disponibilizada pelo Twitter é limitada, não permitindo obter muitos dados em uma única busca e não permitindo também coletar *tweets* mais antigos. Assim, a intenção aqui não foi desenvolver uma aplicação que realiza buscas completamente automáticas, e sim uma aplicação que necessita de parâmetros para poder realizar as buscas. Esses parâmetros são específicos para cada conjunto de buscas, e alguns deles são: descrição do conjunto de buscas, textos buscados *(hashtags* ou nome da série*),* tempo de espera entre uma busca e outra, quantidade de *tweets* coletados por busca e quantidade de buscas que o conjunto terá.

O código fonte do *crawler* está disponível no *GitHub* (https://github.com/alex-olivera/coleta-tweets-tcc).

5.3 Identificação das *hashtags*

Um dos parâmetros mais importantes utilizados pelo *crawler* para realizar as buscas serão as *hashtags*. Elas foram escolhidas a partir da análise de buscas prévias realizadas no Twitter. Alguns cuidados tiveram que ser levados em consideração, como o uso de uma *hashtag* que pode ser usada em outro cenário que não o analisado - um exemplo de problemas que podem acontecer na escolha das *hashtags* acontece com a série Gotham, que não será analisada devido ao período de sua temporada, que requer um cuidado no uso da *hashtag* #gotham, pois as pessoas poderão estar usando essa *hashtag* para outro contexto, como jogos, revistas em quadrinhos e filmes.

5.4 Coleta dos *tweets*

A coleta dos *tweets* iniciou após a conclusão da primeira versão totalmente funcional do *crawler*. A coleta usou as *hashtags* escolhidas previamente e foi realizada em quatro períodos distintos, sendo três deles para as séries lançadas de forma convencional nas emissoras de TV, e um período para as séries disponíveis nos serviços de streaming, sendo uma temporada lançada por completo uma única vez. Para o primeiro caso, os três períodos foram: um dia antes da estreia de um episódio, no dia da estreia do episódio, um dia após a estreia do episódio. Aqui a intenção é observar qual a expectativa para o episódio, o que está sendo comentado na exibição do episódio e qual a opinião das pessoas sobre aquele episódio após sua exibição.

5.5 Pré-processamento dos dados

De posse da base de *tweets*, foi realizado um pré-processamento de todos os dados para a retirada de conteúdos julgados irrelevantes para a análise desejada.

Foram retiradas todas as *URLs* presentes nas mensagens, pois foi considerado aqui que as *URLs* não representam nenhum sentimento e também não serão úteis para as demais técnicas de análise de redes sociais. Um segundo filtro aconteceu para remover dos textos menções a perfis do Twitter, os quais são caracterizados por serem antecedidos pelo caractere @ (arroba).

Logo em seguida, houve a remoção dos *stopwords* presentes nos textos. Para isso, foi usado um conjunto de palavras caracterizadas como *stopwords* do português brasileiro (como artigos e preposições).

Por fim, foi realizada nos textos a aplicação de *steaming*.

5.6 Nuvens de palavras

Foram geradas nuvens de palavras para os *tweets* das séries que serão analisadas. Nuvens de palavras são imagens compostas por palavras, que demonstram de maneira visual a ocorrência de palavras em um texto. Quanto maior a ocorrência de uma palavra, maior será seu tamanho na nuvem. Para a construção das nuvens, foram consideradas séries com mais de 10 mil *tweets*.

5.7 Construção da rede social

Foi gerado um grafo que representa a rede social criada, no qual ele é não direcional. Os nós são as séries, as arestas são as ligações entre duas séries, representando uma relação entre elas, e o peso da aresta foi definida de acordo com o peso da relação.

5.8 Análise de redes sociais

Foram escolhidas cinco métricas para realizar a análise da rede criada. As métricas são: grau do nó, densidade, grau de centralidade, popularidade e grau de proximidade.

O grau do nó foi calculado contando quantos nós um nó está diretamente ligado.

A densidade foi medida através das relações que uma série possui com outra série. Por exemplo, se em uma rede social temos 10 nós, e o nó A possui 7 relações das 9 possíveis, então sua densidade será de 77,77%.

O grau de centralidade foi medido levando em consideração os pesos do relacionamento, ou seja, quanto maior for o peso nos relacionamentos de uma série e quanto mais conexões ela possuir com as demais séries, mais central ela estará.

A popularidade foi medida pela quantidade de comentários que uma série possui, independente da polaridade das mensagens.

Por fim, o grau de proximidade foi medido calculando todas as distâncias geodésicas de um nó para se ligar aos restantes. Cada nó possui um valor de distância geodésica para cada um dos demais nós. A soma desses valores é chamada de distância. De acordo com Alejandro (2005) a proximidade é calculada dividindo o valor de “1” pela distância e multiplicando o resultado por 1000. Quanto maior o valor de proximidade, melhor é a capacidade de um nó se ligar aos demais.

Foi gerada uma tabela apresentando o grau do nó, a popularidade, a densidade, o grau de centralidade e o grau de proximidade das séries. Foi criada uma tabela que apresenta as relações que mais acontecem.

5.9 Mineração de textos

Após ser realizada a etapa de pré-processamento, foi aplicado o algoritmo de classificação de textos Naive Bayes do *RapidMiner* para gerar o modelo de classificação de *tweets* da série de TV *Game of Thrones*.

Para isso, uma pequena quantidade de textos (textos estes já pré-processados) dos *tweets* coletados foi separada para gerar o modelo de classificação. Inicialmente ocorreu uma classificação manual neste conjunto. Logo após, o conjunto foi dividido em dois: conjunto de treino e conjunto de testes. A partir do conjunto de treino, foi criado um modelo de classificação de textos, que agrega as palavras às categorias em que elas acontecem. As categorias no presente trabalho são os sentimentos que um texto pode informar, sendo eles positivo, negativo e neutro. Esse passo é importante, pois a partir dele o algoritmo aprende como associar um sentimento a um texto. Logo após, o modelo de classificação foi validado com o conjunto de treino utilizando o conjunto de testes. A função desta validação é verificar quão bom o modelo obtido é. Por fim, uma amostra de *tweets* foi classificada e analisada pelo autor desde trabalho, com o objetivo de compatibilizar o resultado da acurácia da classificação dos *tweets* da amostra com o resultado da acurácia da classificação do conjunto de teste.

5.10 Análise de sentimentos

Com o modelo de classificação validado, foi possível realizar a análise de sentimentos, classificando os *tweets* de acordo com suas polaridades.

5.10.1 Classificação

Utilizando o modelo de classificação criado na etapa anterior, foi analisada a polaridade dos *tweets* da série *Game of Thrones*, os quais foram categorizados em positivo, negativo ou neutro. A escolha desta série se deu a sua grande popularidade e por apresentar episódios semanais, possibilitando assim uma análise por episódio. Após a classificação, foi realizada uma correlação entre as notas dadas por episódios para a série no site IMDB[4] (*Internet Movie Database*) com as avaliações da classe positiva. A correlação é uma medida padronizada da relação entre duas variáveis que possui variação de -1 a 1, em que -1 é uma perfeita correlação negativa, ou seja, a covariação é inversamente proporcional entre as variáveis, e 1 é uma perfeita correlação positiva, ou seja, existe uma covariação entre as variáveis diretamente

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

proporcional. O IMDB é um dos mais respeitados sites de crítica popular de filmes e séries do mundo[5]. A classificação foi então validada com base nas notas dadas para a série no site.

[5] http://blogs.estadao.com.br/link/imdb-agora-tambem-em-portugues/

6 RESULTADOS

Neste capítulo, serão apresentados os resultados encontrados no trabalho. Foram coletados 708.760 *tweets* distintos referentes a quatorze séries. Foram criadas nuvens de palavras para as séries que apresentaram mais de 20 mil *tweets* coletados. Foi realizada a análise de sentimentos na série *Game of Thrones*. Por fim, foi realizada a aplicação das técnicas de análise de redes sociais para as séries que possuíram mais de 5 mil *tweets* coletados.

6.1 Coleta e pré-processamento dos *tweets*

A coleta foi realizada durante o período de abril a agosto de 2015, iniciando pela série *Demolidor* e finalizando pela série *Narcos*.

Foi observado pelo autor deste trabalho, que a quantidade de *tweets* publicados sobre um episódio exibido de forma semanal era bem maior um dia antes da estreia do episódio, no dia da estreia do episódio e um dia depois da estreia do episódio. Já para as séries em que os episódios são disponibilizados todos de uma única vez, foi observado que a quantidade de *tweets* publicados referentes a série era bem maior durante o primeiro mês de exibição da temporada. Essas duas observações foram importantes para definir quais dias seriam realizadas as coletas.

Para realizar a busca por *tweets* foi necessário fornecer alguns parâmetros de busca para o *crawler*, que são a descrição da busca, o que seria buscado, a quantidade de requisições que seriam feitas na busca, a quantidade de *tweets* que seriam retornados por requisição e o intervalo de tempo entre uma requisição e outra. Para o parâmetro do que seria buscado, foi realizada uma análise prévia no Twitter para identificar quais eram as palavras chaves mais utilizadas nos *tweets* de cada série. Foram utilizadas as seguintes palavras chaves, em que algumas são palavras de *hashtags* com a retirada do caractere #, pertencentes as séries:

- *Game of Thrones*: "Game of Thrones", GameofThrones, GOT, GOT5 e GOTHBO;
- *Orphan Black*: "Orphan Black" e OrphanBlack;
- *Penny Dreadful*: pennydreadful, "penny dreadful" OR SHO_Penny OR Dreadful OR PennyDbrasil;
- *Demolidor*: demolidor, daredevil, demolidornetflix e daredevilnetflix;
- *Orange is the New Black*: "Orange is the New Black", OrangeistheNewBlack, OrangeistheNewBlackNetflix, orangeTBT, OITNB e OrangeCon;
- *Sense 8*: "Sense 8", Sense8 e Sense8netflix;
- *Hannibal*: hannibal e NBCHannibal;

- *True Detective*: "True Detective", TrueDetective, TrueDetectiveHBO, TDHBO;
- *Under the Dome*: "Under the Dome", UndertheDomeTNT e UndertheDome.
- *Narcos*: Narcos e NarcosNetflix;
- *Suits*: Suits, Suits_USA e suitsSpace;
- *The Strain*: "The Strain", TheStrainFX, strainFX e TheStrain;
- *Between*: "between netflix", betweennetflix e between;
- *Wet Hot American Summer*: WHASNetflix, WetHotAmericanSummer e WetHotAmericanSummerNetflix.

Foram coletados ao todo 708.760 *tweets* distintos para todas as séries, em que cada um trazia como atributos o seu id, o seu texto, a data em que foi publicado e o id, o nome e a localidade do usuário que publicou. Os *tweets* coletados foram salvos inicialmente em arquivos de formato csv (formato de arquivo de texto que pode ser usado para trocar dados de uma planilha entre aplicativos) com valores separados pelo caractere ";". Cada arquivo foi separado por diretórios, onde cada diretório pertencia a uma busca realizada.

Após o processo de coleta ter sido encerrado, ocorreu o pré-processamento dos *tweets,* em que informações desnecessárias foram retiradas do texto (links, menções a perfis de usuários, *stopwords*, entre outros já citados anteriormente). Após a finalização do pré-processamento, visando facilitar a construção das buscas necessárias no conjunto da coleta para as análises a serem realizadas posteriormente, os *tweets* foram armazenados em banco de dados. Pelo baixo número de *tweets* coletados, menos de 5 mil, as séries *Wet Hot American Summer*, *Between* e *The Strain* foram descartadas para as análises. A Tabela 3 apresenta as informações sobre as séries coletadas.

Tabela 3 – Informações sobre as séries coletadas

Séries pesquisadas	**Data de início da coleta**	**Número de *tweets* coletados**
Demolidor	09 de abril de 2015	45.895
Game of Thrones	10 de abril de 2015	164.737
Orphan Black	18 de abril de 2015	22.682
Penny Dreadful	03 de maio de 2015	9.198
Between	20 de maio de 2015	1.113
Hannibal	05 de junho de 2015	18.483
Sense 8	05 de junho de 2015	112.182
Orange is the New Black	12 de junho de 2015	192.236

Under the Dome	27 de junho de 2015	18.820
Suits	27 de junho de 2015	5.659
True Detective	22 de junho de 2015	18.071
The Strain	11 de julho de 2015	737
Wet Hot American Summer	31 de julho de 2015	1.956
Narcos	31 de agosto de 2015	96.991

Fonte: elaborada pelo autor

6.2 Nuvens de palavras

Foram criadas nuvens de palavras para as séries coletadas que possuíam mais de 40 mil tweets presentes na base de dados. A intenção aqui é representar graficamente quais as *hashtags* são mais mencionadas nos *tweets* durante uma temporada de uma série. Foi escolhido um desenho personalizado para a nuvem de palavras, representando uma TV.

A primeira nuvem de palavras gerada foi a da série *Demolidor*, como pode ser visto na Figura 3. É possível observar que as menções mais presentes pertencem ao nome da série em português, "demolidor", e o nome da série em inglês, "*daredevil*". As menções aos perfis de usuário do Twitter do @g1 e @bancodeseries são bastante citadas, devido aos seus vários *tweets* publicados e rettwitados. Outras menções que merecem destaque são as palavras "netflix", serviço de *streaming* responsável pela exibição da série, "arrow", série que também é relacionada a herói, e "marvel", empresa criadora do personagem Demolidor.

Figura 3: Nuvem de palavras da série *Demolidor*

Fonte: elaborada pelo autor

A próxima nuvem de palavras, representada na Figura 4 é referente a série *Game of Thrones.* As maiores menções estão associadas ao nome da série, que são: "gameofthrones", "got", "gotbr", "gotseason5" e "gameofthonesseason5", no qual a palavra “got” é uma palavra formada pelas iniciais do nome da série. Outras menções importantes a serem destacadas são as palavras "gotfinale" e "gameofthnonesfinale", referentes ao último episódio da temporada. Os perfis do Twitter @g1 e @bancodeseries são novamente bastante citados. Por não conter apenas um personagem principal, era esperado que alguns se destacassem nas mensagens. Isso ocorreu com os personagens *Jon Snow* e *Daenerys Targaryen* (também chamada durante a série pelo nome *Khaleesi*) que foram bastante citados, representados pelas palavras "jonsnow", "daenerys" e "khaleesi". Uma reclamação constante dos espectadores de uma série, o *spoiler,* foi outra palavra bastante citada, em forma de uma reclamação e em forma de um aviso *(spoiler* é quando alguém disponibiliza uma informação sobre o conteúdo de algum filme, livro ou série, sem que a pessoa que recebeu essa informação tenha visto o filme, livro ou série*).* Novamente o responsável pela exibição da série é bastante citado, com as palavras "hbo" e "hbobr". Por fim, outra palavra bastante citada foi "domingot", que é a junção das palavras domingo e got (a série tem exibição dos seus episódios no domingo).

Figura 4: Nuvem de palavras da série *Game of Thrones*

Fonte: elaborada pelo autor

Podemos analisar a nuvem de palavras da série *Orange is the New Black* na Figura 5. As maiores menções de palavras se referem a variações da escrita do nome da série, que são: "orangeisthenewblack", "oitnb" e "oitnbseason3". Novamente o perfil do Twitter @bancodeseries é bastante citado com a palavra "bancodeseries". O serviço de *streaming*

Netflix, responsável pela exibição da série, é outra palavra consideravelmente citada. Outras palavras que merecem destaque são: "orangecon", palavra referente a um evento que acontece sobre a série, "maratona", referente ao ato de assistir vários episódios em seguida, e "lovewins", *hashtag* bastante usada referente a legalização do casamento homossexual nos Estados Unidos[6].

Figura 5: Nuvem de palavras da série *Orange is the New Black*

Fonte: elaborada pelo autor

A Figura 6 representa a nuvem de palavras da série *Sense 8*. A palavra "sense8", referente ao nome da série, é a mais citada. A palavra "maratona" também é bastante citada neste conjunto, em que se refere ao ato de assistir vários episódios em seguida. Novamente o perfil do Twitter @bancodeseries e @g1 foram bastante citados. A palavra "netflix" foi outra que novamente apareceu com muitas citações (a Netflix é a responsável pela exibição da série). Outras duas palavras merecem destaque, que são a "lovewins", referente a legalização do casamento homossexual nos Estados Unidos, e a palavra "wewantsense8season2", palavra esta que significava uma reinvindicação por parte dos usuários pela renovação para uma próxima temporada da série.

[6] http://noticias.terra.com.br/mundo/estados-unidos/trending-topic-lovewins-internautas-comentam-casamento-gay-nos-eua,448943e155bc4620e6070e2e9bdfd76dmn5uRCRD.html

Figura 6: Nuvem de palavras da série *Sense 8*

Fonte: elaborada pelo autor

Por fim, a Figura 7 representa a nuvem de palavras da série *Narcos*. A palavra mais citada é "narcos", nome da série. Novamente, como acontece em todas as outras nuvens de palavras, o perfil do Twitter @bancodeseries e a Netflix, que é a responsável pela exibição da série, são bastante citados, com as palavras "bancodeseries", "netflix" e "narcosnetflix". O nome do ator Wagner Moura, que interpreta o personagem principal da série aparece com bastante frequência, com a palavra "wagnermoura". Outras três palavras que merecem destaque por seu número de citações são: a palavra "maratona", referente ao ato de assistir vários episódios em seguida, a palavra "padilha", que se refere ao diretor da série José Carlos Padilha, e a palavra "pabloescobar", que se refere a Pablo Escobar, personagem principal da série.

Figura 7: Nuvem de palavras da série Narcos

Fonte: elaborada pelo autor

Alguns dos resultados encontrados eram esperados, entre eles:

- Os nomes das séries e suas várias formas de escrita em uma *hashtag* foram as palavras mais citadas;
- Os perfis de usuários do Twitter que postam notícias sobre as séries também foram bastante citados. Isso aconteceu em especial com dois perfis, o do @g1, famoso site de notícias do Brasil[7] e do @bancodeseries, site que funciona como um organizador de séries que uma pessoa assiste, em que a pessoa pode dar notas aos episódios e receber informações sobre os próximos episódios da série que ela votou[8];
- Referências às emissoras que transmitem as séries, o que aconteceu com citações como "netflix" e "hbo";
- Eventos que ocorreram durante a transmissão da série e tinham alguma relação com ela, caso da aprovação do casamento homossexual nos Estados Unidos, em que a palavra "lovewins" foi bastante citada nas séries *Orange is the New Black* e *Sense 8*, séries estas que tratam como um dos seus temas o homossexualismo;
- Nomes de personagens e atores, como por exemplo, citações ao ator *Wagner Moura* na série Narcos e ao personagem Jon Snow na série *Game of Thrones*.

Com as nuvens de palavras, pôde ser percebido quais as palavras eram mais citadas, podendo em muitos dos casos associar um sentimento a elas. Isso serviu de ajuda para a construção do modelo de classificação, auxiliando na classificação manual realizada para a construção dos conjuntos de treino e teste.

6.3 Construção da rede social

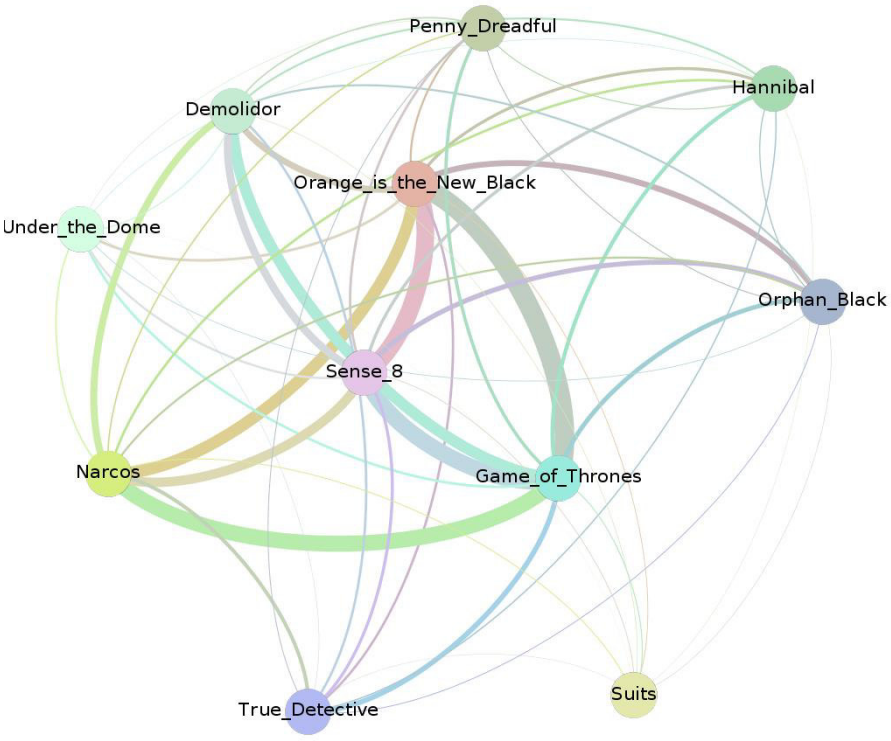
Para a criação da rede social, foram consideradas as séries que possuem mais de 5 mil *tweets.* Os nós da rede foram representados pelas séries. As arestas foram as ligações entre duas séries, ou seja, representa uma relação entre elas. Os relacionamentos foram definidos a partir dos usuários que publicam os *tweets* e as próprias mensagens, em que se duas ou mais séries são comentadas por um mesmo usuário em uma mesma mensagem, elas possuem um relacionamento. O peso de uma relação se deu a partir da quantidade de usuários que comentam

[7] http://g1.globo.com/
[8] http://bancodeseries.com.br/

sobre as séries relacionadas em um mesmo *tweet*, ou seja, se três usuários comentam sobre a série A e a série B, e outros dois usuários comentam sobre a série A e a série C, o relacionamento entre a série A e a série B possuirá peso 3 e o relacionamento entre a série A e a série C possuira peso 2.

A rede criada é não-direcional, possuindo um total de 11 nós e 54 arestas. As arestas são representadas por linhas que realizam a ligação entre dois nós. Quanto maior for a largura da linha, maior o peso do relacionamento entre dois nós. A Figura 8 ilustra o grafo gerado para a rede.

Figura 8: Grafo que representa a rede criada para as séries



Fonte: elaborada pelo autor

6.4 Análise de redes sociais

É possível observar que todos os nós possuem relacionamentos com os demais nós, exceto os nós que representam as séries *Suits* e Penny Dreadful, as quais possuem 9 relacionamentos dos 10 possíveis. A série que obteve maior popularidade foi *Orange is the New Black*, com 192.236 *tweets* publicados. A série que apresentou maior centralidade, foi *Game of Thrones*, possuindo 10 relações das 10 possíveis e um peso total dos seus relacionamentos igual

a 27.710, fazendo assim com que o seu grau de centralidade tenha o valor de 27.720. Das cinco séries com maior popularidade, três delas eram séries estreantes, ou seja, a temporada analisada foi a primeira temporada. Era esperado que estas séries possuíssem muitos *tweets*, haja vista a expectativa que havia pela estreia delas. Outro fator que é possível observar é que, novamente, dentre as cinco séries com maior popularidade, quatro delas possuem os episódios de suas temporadas liberados todos de uma única vez. A Tabela 4 mostra as medidas das métricas analisadas sobre a rede.

Tabela 4: Informações sobre a rede social criada

Série	Grau dos nós	Popularidade	Centralidade	Proximidade	Densidade
Orange is the New Black	10	192.236	24.068	100	100%
Game of Thrones	10	164.737	27.720	100	100%
Sense 8	10	112.182	22.591	100	100%
Narcos	10	96.991	18.781	100	100%
Demolidor	10	45.895	13.971	100	100%
Orphan Black	10	22.682	7.817	100	100%
Under the Dome	10	18.820	4.254	100	100%
Hannibal	10	18.483	6.321	100	100%
True Detective	10	18.071	6.869	100	100%
Penny Dreadful	9	9.198	4.756	90	90%
Suits	9	5.659	1.816	90	90%

Fonte: elaborada pelo autor

A Tabela 5 mostra as quinze relações mais fortes na rede. A série *Game of Thones* é a que possui mais relações fortes, possuindo sete relações dentre as quinze mais fortes. É possível observar que séries da mesma emissora têm relações fortes, em que dentro das quinze relações mais fortes, seis são de duas séries da Netflix, e uma é de duas séries da HBO[9] – apenas

[9] www.hbo.com

duas séries da HBO foram analisadas no presente trabalho. O fato da Netflix oferecer uma grande quantidade de conteúdos disponíveis, no qual suas principais autorias são séries, e também por oferecer recomendações personalizadas para os usuários[10], ela influencia nessas relações. Já no caso da HBO, as duas séries que apresentam uma relação possuem a exibição dos seus episódios no mesmo dia e mesmo horário, transmitidas em épocas diferentes. Porém, logo que a temporada de *Game of Thrones* se encerrou, na semana seguinte deu-se início à temporada de *True Detective*. É possível observar uma continuidade de audiência para as séries da HBO transmitidas nesse horário. Outra observação que pode ser feita é que séries que são transmitidas na mesma época têm relação forte.

Tabela 5: Relações com mais peso dentro da rede social criada

Série 1	Série 2	Peso da relação
Orange is the New Black	*Game of Thrones*	6.507
Orange is the New Black	*Sense 8*	5.888
Game of Thrones	*Sense 8*	5.429
Game of Thrones	*Narcos*	5.242
Orange is the New Black	*Narcos*	3.948
Game of Thrones	*Demolidor*	3.731
Sense 8	*Narcos*	3.153
Sense 8	*Demolidor*	2.556
Narcos	*Demolidor*	2.345
Orange is the New Black	*Demolidor*	2.088
Orange is the New Black	*Orphan Black*	1.952
Game of Thrones	*True Detective*	1.703
Sense 8	*Orphan Black*	1.630
Game of Thrones	*Orphan Black*	1.492
Game of Thrones	*Hannibal*	1.288

Fonte: elaborada pelo autor.

Foi identificado com a pesquisa que usuários que assistem *Game of Thrones* costumam assistir séries de diferentes estilos. Um exemplo disso é a relação que a série possui com *Narcos*. As duas possuem uma relação forte, a quarta maior, porém não possuem nenhuma das características que as relações fortes apresentaram no presente trabalho. Por fim, foi possível observar que os usuários costumam assistir mais de uma série na semana, haja vista o número de relações que acontecem entre duas séries que são transmitidas na mesma época.

Foram observadas a centralidade e a popularidade para a descoberta de qual série seria aplicada a análise de sentimentos. A série escolhida foi *Game of Thrones*.

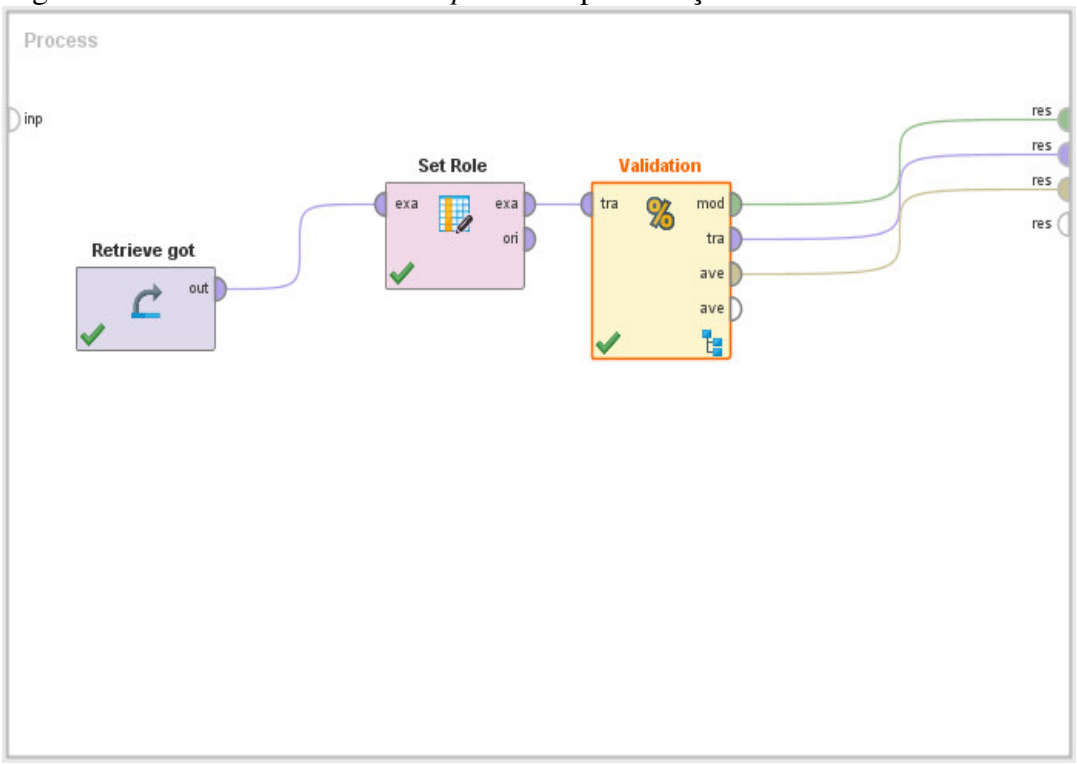
[10] http://www.tecmundo.com.br/netflix/43773-como-a-netflix-faz-para-sugerir-os-filmes-que-voce-quer-ver-.htm

6.5 Desenvolvimento e validação do modelo de classificação

O próximo passo para a realização do trabalho foi o desenvolvimento e validação do modelo de classificação.

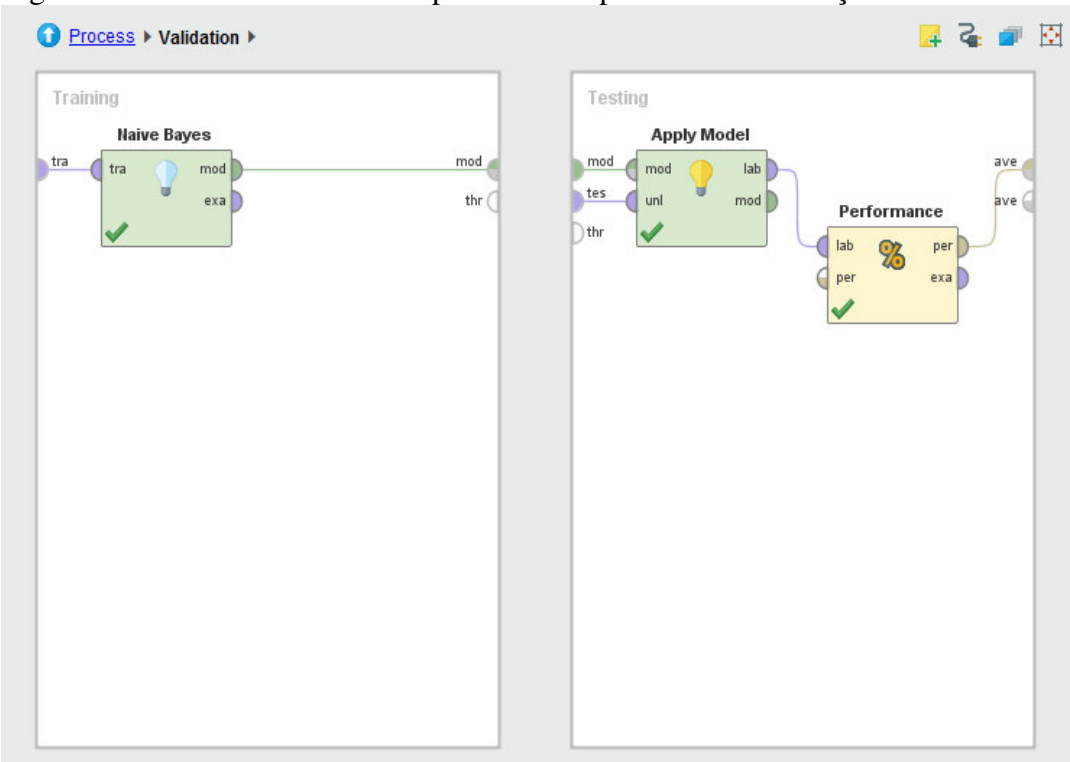
Foi identificado em trabalhos relacionados, que o número de *tweets* para a criação de um bom modelo de classificação varia entre 2 mil e 5 mil mensagens (CARVALHO FILHO, 2014; VIANA, 2014; FRANÇA; OLIVEIRA, 2014). Assim, 3.400 *tweets* relacionados à série *Game of Thrones* foram selecionados de forma aleatória para gerar o modelo de classificação. Esses *tweets* foram classificados de forma manual pelo autor desse trabalho, de acordo com uma das três polaridades definidas (positivo, negativo e neutro). Logo após, o conjunto de treino foi dividido em duas partes, em que 80% (2.720 *tweets*) foram selecionados randomicamente para treinar o algoritmo de Naive Bayes e 20% (680 *tweets*) foram selecionados randomicamente para testar o modelo. A Figura 9 e a Figura 10 mostram o processo realizado no *RapidMiner* para criação e validação do modelo.

Figura 9: Processo realizado no *RapidMiner* para criação do modelo



Fonte: elaborada pelo autor.

Figura 10: Fases de treino e teste presentes no processo de validação



Fonte: elaborada pelo autor.

Após o modelo gerado, foi testada sua acurácia. O modelo obteve uma taxa de acurácia de 73.24%. A Tabela 6 apresenta para as classes classificadas as seguintes informações: precisão, *recall* e *F-score*.

Tabela 6: Precisão, *recall* e *F-score* do classificador para as categorias testadas

	Precisão	***Recall***	***F-score***
***Corpus* de teste Positivo**	58,87%	100%	72,50%
***Corpus* de teste Negativo**	100%	45,50%	62,54%
***Corpus* de teste Neutro**	100%	86,25%	92,61%

Fonte: elaborada pelo autor.

O modelo apresentou-se bom para classificar *tweets* positivos e neutros, já para *tweets* negativos a taxa de erro foi maior.

Por fim, foi submetida uma nova amostra, com apenas 350 *tweets* para serem classificadas com o modelo gerado, em que 200 *tweets* são positivos, 75 *tweets* negativos e 75 *tweets* neutros. Para a classe positiva, o modelo classificou corretamente 84% dos *tweets* (238 instâncias). Para a classe negativa, o modelo classificou corretamente 60% dos *tweets* (45 instâncias). Para a classe neutro, o modelo classificou corretamente 89% dos *tweets* (67

instâncias). No geral, o modelo classificou corretamente 77,66% dos *tweets* pertencentes à amostra. A análise dos dados obtidos com a aplicação do modelo pode ser vista em 6.4.1.

6.6 Análise de sentimentos

Após o modelo de classificação ser validado, o próximo passo foi a classificação do restante dos *tweets* referentes à série *Game of Thrones*, utilizando o modelo de classificação. Foram analisados dez períodos distintos, correspondentes aos episódios. A Tabela 7 apresenta informações sobre cada um dos períodos analisados, contendo os dias em que foram exibidos episódios da série e quais os dias em que foram realizadas coletas para aquele episódio.

Tabela 7: Informações sobre os períodos de análise da série *Game of Thrones*

Número do episódio	Data de exibição	Período da coleta
Episódio 01	12 de abril de 2015	11, 12 e 13 de abril de 2015
Episódio 02	19 de abril de 2015	18, 19 e 20 de abril de 2015
Episódio 03	26 de abril de 2015	25, 26 e 27 de abril de 2015
Episódio 04	03 de maio de 2015	02, 03 e 04 de maio de 2015
Episódio 05	10 de maio de 2015	09, 10 e 11 de maio de 2015
Episódio 06	17 de maio de 2015	16, 17 e 18 de maio de 2015
Episódio 07	24 de maio de 2015	23, 24 e 25 de maio de 2015
Episódio 08	31 de maio de 2015	30 e 31 de maio e 01 de junho de 2015
Episódio 09	07 de junho de 2015	06, 07 e 08 de junho de 2015
Episódio 10	14 de junho de 2015	13, 14 e 15 de junho de 2015

Fonte: elaborada pelo autor

6.6.1 Classificação

A partir do modelo de classificação criado, foram classificados todos os *tweets* coletados e pré-processados pertencentes à série *Game of Thrones*, em um total de 164.737 *tweets*. A Tabela 8 apresenta os resultados obtidos durante toda a exibição da temporada de 2015, também apresentando a nota por episódio fornecida no site IMDB, junto com o total de avaliações que ocorreram.

Tabela 8: Classificação dos *tweets* da série *Game of Thrones*

Episódio	Qtd. tweets	Positivo	Negativo	Neutro	IMDB
Episódio 01	18.209	90,8%	3%	6,2%	8.4/10 (11.195)
Episódio 02	14.039	89,4%	4,08%	6,52%	8.5/10 (8.229)
Episódio 03	3.081	88,32%	4,5%	7,18%	8.5/10 (7.600)
Episódio 04	11.887	89,88%	3,17%	6,95%	8.7/10 (8.137)
Episódio 05	14.465	89,21%	4,3%	6,49%	8.6/10 (8.358)
Episódio 06	12.898	88,21%	6,2%	5,59%	7.8/10 (9.536)
Episódio 07	12.349	94,5%	2,29%	3,21%	9.0/10 (9.962)
Episódio 08	7.416	96,49%	1,2%	2,31%	9.9/10 (39.659)
Episódio 09	17.426	91,5%	3,21%	5,29%	9.4/10 (17.986)
Episódio 10	15.663	90,21%	4,2%	5,59%	8.6/10 (18.183)

Fonte: elaborada pelo autor

É possível observar que o número de classificações positivas é bem superior às demais. Alguns episódios se destacam pela sua classificação positiva acima dos 90%. No primeiro episódio muitos *tweets* eram de euforia pela volta da série. Um exemplo disto são os *tweets* "5° temporada de GAME OF THRONES até q enfim lançou amo essa série, é muito boooooooa ela é d+ http://t.co/4TQsac07aw" e "Ansiedade? Amanhã! Nem acredito! Contando as horas! #GoTSeason5 #GoT http://t.co/i2No2O4bwH". Outros episódios que obtiveram classificação positiva acima de 90% foram os últimos quatro episódios, mostrando um crescente de avaliações positivas no final da temporada.

O número de classificações neutras seguiu em todos os episódios superior ao número de classificações negativas. No geral, as classificações neutras se deram em sua maioria por notícias publicadas como *tweet*. Alguns exemplos disso são os *tweets* "@omelete: De ontem: George R.R. Martin apresenta Lil Thrones, paródia animada de Game of Thrones http://t.co/RyL4sLKS0i http://t.co/GywY…" e "@HBO_Brasil: A estreia de Game of Thrones acontece hoje, às 22h03, no canal HBO".

O destaque nos *tweets* de classificação negativa se dá por três fatores. O primeiro é o fato do usuário não gostar da série. Um *tweet* exemplo é "Me sinto meio deslocada por não gostar de #GameofThrones". O segundo fator aconteceu no início da temporada, com o vazamento de quatro episódios antes de serem transmitidos de forma oficial[11]. Um *tweet*

[11] http://g1.globo.com/pop-arte/noticia/2015/04/quatro-episodios-da-nova-temporada-de-game-thrones-vazam-na-web.html

exemplo é “ja vazaram os 4 primeiros episodios de game of thrones, este povo não presta”. Um reflexo desse fator pode ser observado no baixo número de *tweets* referentes ao terceiro episódio, ou seja, um baixo interesse pela transmissão do episódio pela HBO, no qual grande parte das reclamações em relação ao vazamento aconteceram nos dois primeiros episódios. O terceiro fator é uma reclamação constante durante toda a temporada, que é o *spoiler*. Um exemplo disso é o *tweet* “Twitter e Facebook virados em Spoilers de Game of Thrones... que queimem na chama do inferno essa gente...”.

O resultado da classificação pôde ser validado realizando uma correlação com os dados fornecidos pelo site IMDB. A correlação apresentada foi igual a 0.7713, indicando que existe uma forte correlação positiva entre as duas variáveis.

É possível observar que as notas por episódio fornecidas no site apresentam um crescente com o passar da temporada, assim como aconteceu na classificação do presente trabalho. Outro fator a ser observado é que apenas uma nota fica abaixo de 8, e no presente trabalho todas as avaliações obtiveram classificação positiva superior a 80%. Assim como aconteceu na coleta, o episódio que obteve menos avaliações foi o terceiro.

7 TRABALHOS FUTUROS

Foram aplicadas neste trabalho algumas métricas de análise de redes sociais, que são julgadas pela literatura como as principais. É possível usufruir do uso de mais métricas para realizar novos tipos de análises na rede interativa criada.

Foram coletados *tweets* referentes a quatorze séries de TV. Para uma análise mais ampla é possível a coleta para um número maior de séries, aumentando assim a possibilidade de relações criadas e melhorando algumas análises que já foram feitas no trabalho.

A rede social criada foi baseada na relação que as séries possuíam, em que as séries eram os nós e as arestas eram as relações entre duas delas. Para a construção de novas avaliações, é possível criar a rede social com os usuários e as séries sendo os nós, e as arestas sendo as relações de um usuário com uma série. Esse tipo de rede possibilita a aplicação de técnicas de recomendação de conteúdo.

Em relação a análise de sentimentos, foi aplicado neste trabalho a implementação do algoritmo de classificação de textos Naive Bayes, disponibilizada pelo *RapidMiner*. Outras implementações e algoritmos de classificação podem ser utilizadas e comparadas.

Para obtenção de melhores resultados na classificação, é possível aperfeiçoar o modelo de classificação, utilizando um conjunto maior de *tweets* e de forma mais balanceada entre as classes. Outras categorias podem ser definidas, de acordo com o contexto do trabalho.

A classificação foi aplicada a apenas uma série. Para ampliar os resultados, é possível avaliar mais séries.

A rede social utilizada neste trabalho foi o Twitter. Ainda que seja definido pela literatura como a melhor rede social para aplicar análise de opiniões, as mensagens podem ser extraídas de outras fontes, como o Facebook.

8 CONCLUSÃO

Este trabalho teve como objetivo coletar e analisar mensagens da rede social Twitter. A coleta foi realizada entre os meses de abril e setembro de 2015, em que foram coletados no total *tweets* referentes a quatorze séries de TV, as quais apresentam duas formas de exibição dos seus episódios: exibidos semanalmente e a temporada liberada toda de uma única vez por um serviço de *streaming*. Algumas séries foram descartadas para a fase de análise por seus baixos números de *tweets* coletados.

Para a análise foram utilizadas métricas de ARS e classificação por análise de sentimentos. Foram utilizadas as métricas julgadas mais importantes pela literatura. A rede social criada mostrou que a grande maioria das séries coletadas possui relação, ou seja, existem dentro da base criada usuários que comentam sobre mais de uma série. Foi identificado que as relações mais fortes acontecem entre séries de mesma emissora e séries que são transmitidas na mesma época. Foi identificado que, dentre os *tweets* coletados, a série *Orange is the New Black* foi a que apresentou maior número de mensagens. Por fim, foi identificado que a série que possui maior número de relações com outras é *Game of Thrones.*

Devido ao grande tempo gasto com o pré-processamento das mensagens e a criação de um bom modelo de classificação, apenas uma série pôde ser avaliada aplicando a análise de sentimentos. *Game of Thrones* foi a série escolhida para ser aplicada a análise. No total, foram avaliados dez momentos distintos, ou seja, dez episódios. O número de *tweets* positivos ocorreu de forma bem mais ampla do que os demais. Isso se deu pela grande popularidade e aceitação da série pelos usuários[12]. Todos os episódios obtiveram classificações positivas acima dos 80%, em que quatro delas foram acima dos 90%. O resultado pôde ser validado comparando a classificação com informações sobre avaliações da série disponíveis no site IMDB.

Foram criadas nuvens de palavras para as séries que possuíam um total de *tweets* coletados superior a 20.000 mensagens, que permitiram a identificação das palavras mais citadas nas séries.

Assim, o processo aplicado neste trabalho, pode ser seguido por organizações para mapear dentro do Twitter relações entre usuários ou produtos, verificação de que palavras estão sendo mais comentadas e também as opiniões dos usuários, fazendo o uso dos resultados para os mais diversos fins.

[12] http://diversao.terra.com.br/arte-e-cultura/com-recorde-de-indicacoes-game-of-thrones-e-a-serie-a-ser-batida-no-emmy,8670b01ce94cc1ca718e341e8e7b004eh3dqbot0.html

REFERÊNCIAS

ALEJANDRO, V. A.; NORMAN, Aguilar G. **Manual introdutório à análise de redes sociais**. UAEM - Universidad Autonoma Del Estado de Mexico, 2005.

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional**. 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontífica Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2007.

CARVALHO FILHO, José Adail. **Mineração de textos: análise de sentimentos utilizando tweets referentes à Copa do Mundo 2014**. 2014. 44 f. TCC (graduação em Engenharia de Software) - Universidade Federal do Ceará, Campus Quixadá, Quixadá, 2014. Disponível em: <http://www.repositoriobib.ufc.br/000017/0000179f.pdf>. Acesso em: 10 fev. 2015.

FRANÇA, T. C.; OLIVEIRA, J. **Análise de Sentimento de tweets relacionados aos protestos que ocorreram no Brasil entre Junho e Agosto de 2013**.In: III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2014, Brasilia. Anais do Congresso da Sociedade Brasileira de Computação, 2014.

GOLBECK, J. A. **Computing and Applying Trust In Web-Based Social Networks**. Ph.D. Thesis, University of Marylan, College Park, MD, USA, 2005.

GOMES, Helder Joaquim Carvalheira. **Text Mining: análise de sentimentos na classificação de notícias**. Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on. Lisboa. 2013.

LI, Y. M.; LI, T. Y. **Deriving Marketing Intelligence over Microblogs**. Proceedings of 44th Hawaii International Conference On System Sciences (HICSS), pp. 1 –10, 2011.

LI, G.; LIU, F. **A clustering-based approach on sentiment analysis**. Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on (pp. 331–337), 2010.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, Maio de 2012. Synthesis Lectures on Human Language Technologies, 2012.

MEIRA, Silvio; COSTA, Ricardo; JUCÁ, Paulyne Matthews **Redes Sociais**. In: Mariano Pimentel; Hugo Fuks. (Org.). Sistemas Colaborativos. 1ed.Rio de Janeiro: Elsevier-Campus-SBC, 2011, v. 1, p. -, 2011.

NAAMAN, Mor; BOASE, Jeffrey; LAI, C. H. **Is it all About Me? User Content in Social Awareness Streams**. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, 2010.

NASCIMENTO, P.; AGUAS, R.; Lima, D.; KONG, X.; OSIEK, B.; XEXÉO, G.; SOUZA, J. **Análise de sentimentos de tweets com foco em notícias**. In: III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012, Curitiba. Anais do Congresso da Sociedade Brasileira de Computação, 2012.

OWEN, Sean et al. **Mahout in Action**. Connecticut: Manning Publications Co, 2011. 373p.

PASSMORE, David L. **Social network analysis: Theory and applications**. 2011 <http://train.ed.psu.edu/WFED-543/SocNet_TheoryApp.pdf>. Acesso em 03 mar. 2015.

RECUERO, Raquel. **Contribuições da Análise de Redes Sociais para o estudo das redes sociais na Internet: o caso da hashtag# Tamojuntodilma e# CalaabocaDilma**. Fronteiras-estudos midiáticos, v. 16, n. 2, p. 60-77, 2014.

RENNIE, J. D. et al. **Tackling the poor assumptions of naive bayes text classifiers**. In: ICML. 2003. p. 616-623.

RUSSEL, Mathew A. **Mining the social web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More**. 2 ed. Sebastopol: O'reilly Media, Inc., 2013.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining: Mineração de Dados**. Rio de Janeiro: Ciência Moderna, 2009. 900 p. Tradução de Introduction to Datamining, 2009.

VIANA, Zarathon Lopes. **Mineração de textos: análise de sentimentos utilizando Tweets referentes às eleições presidenciais 2014**. 2014. 32 f. TCC (graduação em Sistemas de Informação) - Universidade Federal do Ceará, Quixadá, 2014. Disponível em: <http://www.repositoriobib.ufc.br/000017/000017d1.pdf>. Acesso em: 20 fev. 2015