É possível descrever episódios de séries de televisão a partir de comentários online?

Túlio C. Loures[1], **Pedro O.S. Vaz de Melo**[1], **Adriano A. Veloso**[1]

[1]Departamento de Ciência de Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG – Brazil

{loures.tc,olmo,adrianov}@dcc.ufmg.br

Abstract. *Due to the omnipresence of the Internet and the Web 2.0 in current society, it has become easy to find groups or communities of people discussing the most varied subjects. In this paper, we try to answer the question about whether, even when nothing is explicitly known about the entity referred to in the discussion, it is possible to formulate a general and brief idea of its characteristics when reading comments about it. To study this problem, we analyze the potential that comments have to describe television series, and perform a task of comment classification in order to identify which series and episode it is associated with. This classification task serves as the basis for a method that selects comments with high descriptive value for the episodes and series. Results reveal that a small set of comments can describe their episodes and, when taken together, the series as a whole.*

Resumo. *Por causa do uso onipresente da Internet e da Web 2.0 na sociedade atual, é fácil encontrar grupos ou comunidades de pessoas que discutem sobre os mais variados assuntos. Neste artigo, tentamos responder a questão sobre se, mesmo quando nada é explicitamente conhecido sobre a entidade referida na discussão, é possível formular uma ideia geral e resumida de suas características ao ler comentários sobre ela. Para estudar esse problema, analisamos o potencial que comentários têm para descrever séries de televisão, e realizamos uma tarefa de classificação de comentários a fim de identificar a qual série e episódio ele está associado. Essa tarefa de classificação serve como base de um método que seleciona comentários de alto valor descritivo para os episódios e séries. Resultados revelam que um pequeno conjunto de comentários conseguem descrever seus episódios e, quando tomados em conjunto, a série como um todo.*

1. Introdução

A Internet é composta por milhões de dispositivos em que cada um deles é responsável pela geração, armazenamento e transmissão de incontáveis dados. Neste contexto, os algoritmos de aprendizagem automática têm sido amplamente utilizados para processar e extrair informações valiosas de todos esses dados. Para fazer isso, os algoritmos precisam de representações formais das características das entidades que querem aprender sobre, o que normalmente é uma tarefa desafiadora [Bengio et al. 2013, Ji and Eisenstein 2014]. Esta é uma tarefa ainda mais difícil se essas entidades não têm qualquer informação estruturada explicitamente associada a elas. Considere, por exemplo, o problema de descrever

o conteúdo de um vídeo pessoal postado no Facebook ou de um evento associado a uma *hashtag* do Twitter.

À medida que as tecnologias da Web 2.0 incentivam a contribuição dos usuários ao invés de simplesmente oferecer informações, agora mais e mais pessoas podem comentar livremente sobre diferentes tipos de entidades. Assim, neste artigo investigamos a seguinte pergunta: é possível aprender traços que caracterizam entidades usando apenas discussões online associadas a elas como fonte de conhecimento? Uma clara vantagem de usar comentários em vez de informações estruturadas explícitas como fonte de dados é que os comentários são onipresentes na Internet. Em outras palavras, vemos comentários sobre uma determinada entidade como *crowdsourcing*, similar ao que é feito no jogo ESP [von Ahn and Dabbish 2004], um jogo online licenciado pelo Google que é usado para melhorar a precisão do seu mecanismo de buscas de imagens. Com isso, esperamos gerar atributos para uma entidade sem que nunca alguém tenha que explicitamente explicar o que é.

Além disso, por meio de análises exploratórias, este artigo também investiga o quanto pode ser aprendido sobre uma entidade unicamente a partir de conversas gerais sobre ela. Sabe-se que extrair informações relevantes a partir de comentários é uma tarefa muito desafiadora [Hsu et al. 2009, Siersdorfer et al. 2014, Cheng et al. 2015]. Comentários online são geralmente curtos, e é comum usuários fazerem uso de textos informais e não estruturados para se expressarem, através de, por exemplo, siglas e encurtamento de palavras. Outra dificuldade reside no fato de que comentários permitem que pessoas iniciem e permeiem conversas, que muitas vezes são sobre assuntos bem diferentes da entidade em si. Assim, qualquer método de aprendizado de traços sobre entidades a partir de comentários deve ser capaz de desconsiderar (ou filtrar) esse tipo de conversa.

Para estudar esse problema, analisaremos o potencial que comentários têm para descrever séries de televisão. Uma série de televisão é um tipo de programa televisivo com um número pré-definido de emissões por temporada, chamadas episódios. A partir de sequências de comentários postados em fóruns online, investigaremos se os mesmos podem (e como podem) ser usados para gerar representações formais que descrevem a série de televisão associada a eles. Mais especificamente, neste artigo discutiremos as seguintes perguntas:

1. Dada uma sequência de comentários associada a um episódio e um sumário manualmente criado do mesmo, quanto deste sumário tal sequência pode gerar automaticamente?
2. É possível criar uma representação vetorial de comentários que permita algoritmos de aprendizado de máquina classificar a qual série e episódio um dado comentário pertence?
3. Caso seja possível, quantos e quais comentários são suficientes para cobrir e explicar os acontecimentos da série de televisão associada a eles?

O restante deste artigo está organizado da seguinte maneira. Os trabalhos relacionados são descritos na Seção 2 e a descrição do problema, juntamente com a notação usada neste artigo, são descritos na Seção 3. Na Seção 4, descrevemos o conjunto de dados. Uma análise sobre o potencial que comentários têm para explicar séries de televisão é mostrada na Seção 5. Na Seção 6 é descrito um método para gerar representações formais de comentários e, a partir dela, extrair um conjunto de comentários capazes de explicar

os acontecimentos de uma dada série de televisão. Por fim, na Seção 7, descrevemos as conclusões deste trabalho.

2. Trabalhos Relacionados

Nos últimos anos, tem havido um grande número de estudos voltados para a aprendizagem de representação de entidades dos mais variados tipos e para as mais variadas tarefas. Em [Yang et al. 2015, Perozzi et al. 2014], por exemplo, os autores usam informações diferentes encontradas em grafos para aprender representações de suas redes sociais. Krizhevsky et al. [Krizhevsky et al. 2012] usou algoritmos sofisticados para aprender representações de imagens em um grande conjunto de dados. Estes estudos, no entanto, utilizam informações explícitas extraídas da própria entidade para aprender a representá-la. Em [Bickel and Scheffer 2004], informações encontradas fora da página da web (texto âncora dos links de entrada), bem como a própria página, são usadas para criar uma representação para ela. Estas ideias são similares às deste artigo, mas no nosso caso, o uso de comentários como fonte de dados levanta muitos desafios que, individualmente, já servem de inspiração para estudo. Em [Siersdorfer et al. 2014], por exemplo, os autores analisaram comentários postados em diferentes sites de mídia social, estudando principalmente como o seu conteúdo e classificação podem ser relacionados. O trabalho de [Kleinberg 2002], por sua vez, desenvolve uma abordagem formal para a modelagem de "rajadas de atividade", que pode ser aplicado a fluxos de comentários. Choi et al. [Choi et al. 2015] caracterizam conversas coletadas do *Reddit* em termos de volume, capacidade de resposta, e viralidade.

Outra área da literatura intimamente relacionada a este artigo é a área de sumarização de textos. Para o melhor de nosso conhecimento, a maioria dos trabalhos dessa área foca na sumarização de textos completos, como livros e notícias [Liu et al. 2015b, Moratanch and Chitrakala 2016, Gambhir and Gupta 2017], ou sumarização de opiniões a partir de *reviews* feitos por usuários [Ganesan et al. 2010, Potthast and Becker 2010, Ganesan et al. 2012]. Mais recentemente, métodos foram propostos para extrair informações relevantes a partir de comentários. Khabiri et al. [Khabiri et al. 2011] propuseram uma abordagem baseada em clusterização para identificar grupos de comentários correlacionados em vídeos do Youtube. Yang et al. [Yang et al. 2011] apresentam um método para gerar sumários de páginas da Web considerando também os comentários associados a elas e a rede social entre os usuários que comentaram. Em [Kågebäck et al. 2014], os autores propõe o uso de representações vetoriais de sentenças como métrica de similaridade no processo de geração de sumários por extração. Liu et al. [Liu et al. 2015a] modelam o problema de sumarizar comentários como um problema de agrupamento, em que o número de tópicos coberto pelos comentários deve ser conhecido *a priori*.

Este artigo se diferencia dos trabalhos citados nesta seção em três principais aspectos. Primeiro, este é um trabalho que foca na caracterização do potencial de comentários como fonte de dados única para geração de representações formais de entidades a serem usadas por algoritmos de aprendizagem de máquina. Segundo, através dessa caracterização, investigamos e quantificamos as diferenças entre a linguagem usada em resumos formais e manualmente gerados da linguagem usada por usuários em comentários. Por fim, focamos na descoberta de conhecimento relevante em comentários para representar episódios e séries de televisão. No melhor do nosso conhecimento, este é o primeiro artigo que aborda qualquer um desses três aspectos.

3. Definição do Problema

Para o restante deste trabalho, definimos por c um comentário submetido a uma discussão online, composto por um conteúdo textual. Uma sequência de comentários $\mathcal{C}$ é um conjunto ordenado de n comentários $\{c_1, c_2, ..., c_n\}$ retirados de um contexto em comum.

Definimos também um episódio e de uma série de televisão como uma tupla $(r^e, \mathcal{C}^e)$, na qual r^e é o resumo do episódio e $\mathcal{C}^e = \{c_1^e, c_2^e, ..., c_{n^e}^e\}$ a sequência de n^e comentários referente ao tópico de discussão online daquele episódio, daqui em diante também referida simplesmente por "comentários do episódio" ou "tópico do episódio". Uma sequência de episódios $\mathcal{E}$ é um conjunto ordenado de m episódios $\{e_1, e_2, ..., e_m\}$ retirados de um contexto em comum.

Finalmente, estabelecemos uma série de televisão s como uma tupla $(R^s, \mathcal{E}^s)$, onde R^s é o resumo da série como um todo e $\mathcal{E}^s = \{e_1^s, e_2^s, ..., e_{m^s}^s\}$ é o conjunto ordenado dos m^s episódios pertencentes àquela série. Partindo disso, podemos definir também $\mathcal{C}^s$ a sequência de comentários da série como um todo, equivalente à concatenação de todas sequências $\mathcal{C}^{e_i}$ para $i \in \{1, ..., m^s\}$.

Queremos encontrar um conjunto conciso de comentários que seja capaz de descrever bem um contexto. Para este artigo, estaremos focando no caso de sequências de comentários referentes a episódios de séries. Tendo sidas estabelecidas as definições dos conceitos acima, podemos escrever essa tarefa como: encontrar um $\mathcal{C}^x \subseteq \mathcal{C}^s$ de poucos comentários para cada episódio e tal que R^s seja bem explicado por $\mathcal{C}^x$.

4. Base de Dados

Para este trabalho, foram coletados todos os comentários diretamente associados a episódios de 844 séries de animação do portal *MyAnimeList.net*[1]. Além dos dados de comentários, foram também coletados, quando disponível, resumos r^e manualmente criados para cada episódio e. Esses resumos foram obtidos a partir de páginas da Wikipédia dedicadas a listar e descrever episódios de cada série[2]. O resumo R^s de cada série s é simplesmente a concatenação dos resumos de todos os episódios da série em questão.

Como o número de comentários e episódios varia significativamente de série para série, foram utilizados três critérios para selecionar as séries que analisamos neste artigo. Primeiramente, buscou-se escolher aquelas séries com um número de comentários similar, de forma que os resultados não sejam enviesados para uma série com um número extremo de comentários ou que séries com poucos comentários sejam diluídas nos resultados. O segundo fator determinante para a escolha foi a existência de um resumo facilmente acessível para cada episódio da série. Por fim, escolhemos um conjunto de séries com um número de episódios na ordem de dezenas e comentários na ordem de milhares. A Tabela 1 apresenta as séries que foram escolhidas para análise, juntamente com o número de episódios e de comentários associados a elas[3].

Uma vez coletados os textos descritos acima, tanto dos comentários quanto dos resumos, foram realizadas uma sequência de operações sobre eles de forma a transformar

[1] `https://myanimelist.net/`

[2] Um exemplo de uma dessas páginas pode ser encontrado em `https://en.wikipedia.org/wiki/List_of_Mushishi_episodes`.

[3] Cada série pode ser acessada em `https://myanimelist.net/anime/<ID>`, substituindo <ID> pelos valores dados na tabela.

Tabela 1. Número de episódios e número de comentários para cada série da base de dados.

ID da Série	# Episódios	# Comentários
1	26	2675
19	50	3701
30	26	3540
205	26	1812
226	13	1459
356	24	2065
457	26	2833
777	10	1663
790	23	2037
820	50	3630
877	47	2906
934	26	3529
13599	22	3643
Total	369	35493
Média	28,38	2730,23

o texto original em dados trabalháveis. Primeiramente, foram removidas as citações de outros comentários presentes nos textos. Em seguida, a formatação HTML dos textos foi eliminada, deixando apenas o texto do comentário não-formatado. Os caracteres não-imprimíveis foram excluídos. Finalmente, foram removidas as *stop words* dos textos, e todas as letras foram convertidas para suas formas minúsculas.

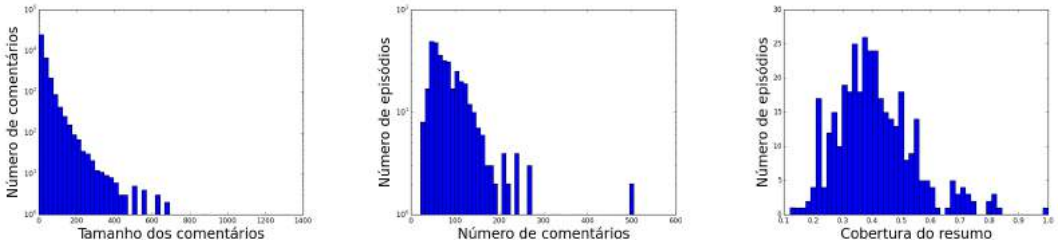
5. Análise e Caracterização de Discussões

Como mencionado na Seção 2, comentários são textos com características bem distintas, podendo variar significativamente em tamanho, forma e conteúdo. Assim, para melhor entender os dados coletados, nesta seção foi realizada uma série de análises sobre as características das sequências de comentários. Tais análises servirão como base para a metodologia apresentada nas seções seguintes.

Na Figura 1(a) é mostrado o histograma para o tamanho (número de termos) dos comentários da nossa base de dados. Observe que a grande maioria dos comentários apresenta um pequeno número de termos, enquanto alguns poucos comentários são bem extensos, possuindo mais de 500 palavras. Tal resultado evidencia a dificuldade de extrair comentários relevantes em sequências de comentários, pois a grande maioria é composta de comentários pequenos, muito provavelmente com pouco conteúdo informativo.

A fim de investigar se há discrepância na participação de usuários por episódio, mostramos na Figura 1(b) o histograma do número de comentários postados por episódio. Observe que muitos episódios possuem um pequeno número de comentários associados a eles, i.e., a maioria dos episódios possui entre 20 e 100 comentários. Semelhante ao que foi exibido na Figure 1(a), há também episódios com uma grande quantidade de comentários. Assim, um método para identificar comentários relevantes e descritivos enfrenta, naturalmente, o problema de ter pouca e excessiva informação disponível ao mesmo tempo.

A Figura 1(a) sugere que boa parte dos comentários não possui informação útil, dado seus tamanhos relativamente pequenos, mas a Figura 1(b) indica que a maioria dos episódios terá uma quantidade suficiente de comentários para que possamos inferir uma descrição dos mesmos. Para chegar a uma forma de determinar quais comen-



(a) Tamanho (b) Comentários por episódios (c) Cobertura por episódio

Figura 1. A Figura 1(a) mostra a distribuição de tamanho do texto (tratado) dos comentários, em escala log. A Figura 1(b) mostra a distribuição do número de comentários por episódio, em escala log. A Figura 1(c) mostra a distribuição da proporção dos resumos das séries cobertos pelos seus respectivos comentários.

tários teriam maior valor descritivo do resumo, primeiro foi feita uma análise de características que pudessem ser usadas para definir e representar os diferentes termos presentes na base de dados. Após testar algumas diferentes opções, observando as representações resultantes por meio do método de redução de dimensionalidade t-SNE[4] [van der Maaten and Hinton 2008], determinamos que o seguinte conjunto de características apresenta informações suficientes e relevantes para a tarefa de determinar a utilidade de um comentário para uma série:

- Contagem total de ocorrências do termo;
- Entropia de Shannon do termo entre as séries;
- Entropia de Shannon do termo entre todos os episódios da base;
- Entropia de Shannon do termo entre os episódios para cada série;
- TF-IDF do termo para cada série;
- Probabilidade do termo ocorrer em cada série.

A Figura 2 mostra a visualização gerada pelo t-SNE para todas as palavras da base representadas por esse conjunto de características. Na imagem, a cor de cada termo representa a série na qual ela apresenta a maior contagem. Observa-se que, apesar da maioria das palavras se misturar em um grande agrupamento, há alguns agrupamentos menores que contêm palavras associadas a uma mesma série (pontos da mesma cor). Uma vez que as características usadas para gerar o T-SNE refletem a presença da palavra nas séries analisadas, foi verificado que esses agrupamentos contêm aquelas poucas palavras que ocorrem predominantemente em uma dada série (nomes de personagens, e outros termos específicos da série). Assim, pode-se concluir que, de fato, o número de palavras que têm o seu uso altamente concentrado em uma única série é pequeno. Isso significa dizer que se um comentário contém uma dessas palavras, provavelmente saberemos dizer a qual série ele está associado. Essa hipótese será explorada nas seções seguintes deste artigo.

Uma outra análise fundamental para entender o potencial de extrair comentários relevantes a partir de comentários é investigar a capacidade que cada comentário tem de explicar o resumo manualmente gerado do episódio no qual ele está associado. Para avaliar isso, foi utilizada uma métrica inspirada na ROUGE-N [yew Lin 2004], para $N = 1$.

[4] `https://lvdmaaten.github.io/tsne/`

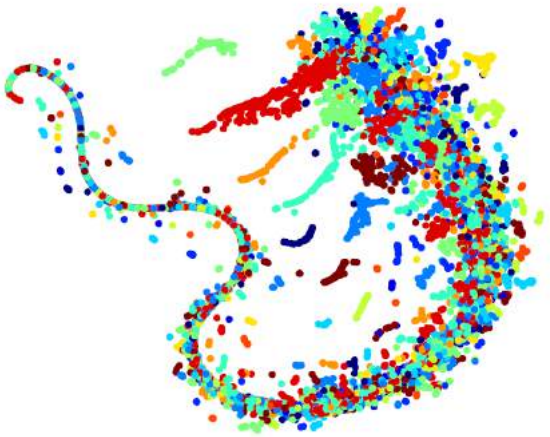


Figura 2. Representação 2D gerada pelo t-SNE das palavras da base, coloridas de acordo com a série em que mais parecem

Essa métrica informa a proporção de palavras do resumo que também estão presentes nos comentários considerados. A Figura 1(c) mostra a distribuição da cobertura dos resumos r^{e_m} de cada episódio e_m pelas suas respectivas sequências de comentários $\mathcal{C}^{e_m}$. Observe que, na média, apenas 40% das palavras do resumo de um episódio são encontradas nos comentários associados a ele.

6. Recuperação de Comentários Representativos

6.1. Representação de comentários

Antes de aplicar métodos que recuperem aqueles comentários que melhor explicam o resumo de uma série ou episódio, precisamos definir uma representação formal (estruturada) para eles. Uma forma simples de se fazer isso é por meio de uma representação estilo *bag-of-words*, i.e., cada comentário c é representado por um vetor binário $\mathbf{w}^c = (w_1^c, ..., w_D^c)$, em que D é o tamanho do dicionário (número de palavras) e $w_j^c = 1$ se a palavra de índice j do dicionário aparece no comentário c, e $w_j^c = 0$ caso contrário.

No entanto, um problema dessa abordagem é a definição do dicionário. No nosso caso, entretanto, pela caracterização feita na Seção 5, sabemos que apenas uma pequena proporção das palavras possuem informação relevante para agrupamento. Portanto, podemos representar os comentários apenas com informação das palavras mais informativas. Assim, estabelecemos um parâmetro K para indicar o número de palavras mais representativas de uma dada série ou episódio, selecionadas a partir da métrica TF-IDF. As palavras obtidas como resultado desse processo são referidas por "*top*-K palavras", e o vetor que representa cada comentário é dado por $\mathbf{w}^c = (w_1^c, ..., w_{K \times N}^c)$, em que N é o número de episódios.

6.2. Seleção de comentários

Através de análises preliminares, observou-se que uma grande parte de comentários não contêm sequer uma palavra entre as *top*-K palavras que discriminam a sua série ou episódio. Tais comentários dificilmente contêm informações relevantes e representativas da série ou episódio ao qual ele está associado. Por isso, implementou-se um segundo parâmetro, α, que indica o número mínimo de palavras relevantes que um comentário c deve

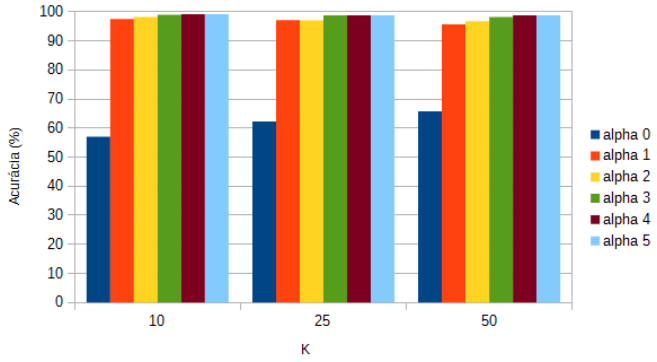


Figura 3. Porcentagens de comentários corretamente classificados em suas respectivas séries, de acordo com valores de K e α.

ter para que ele seja considerado nas próximas etapas, ou seja, $\sum_{j}^{K \times N} w_j^c > \alpha$. Assim, a partir dos parâmetros K e α, desejamos selecionar um subconjunto relevante e descritivo de comentários para a série ou episódio em questão. Aqueles comentários que claramente são associados a sua série ou episódio pela representação proposta são bons candidatos a comentários bem descritivos dos mesmos.

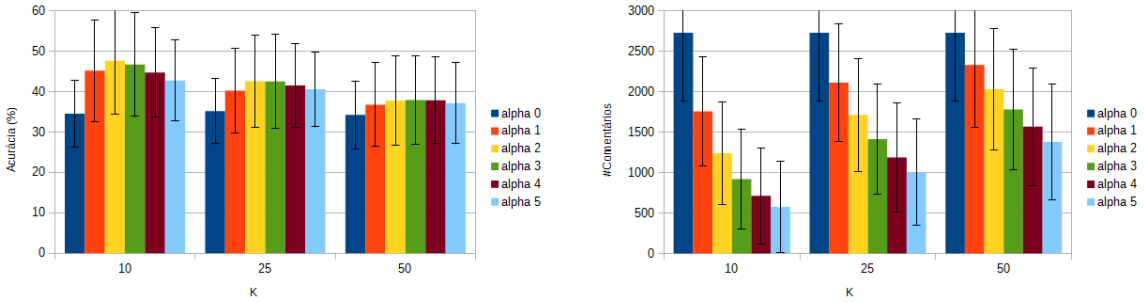
Com este objetivo, foram definidas duas tarefas de classificação para identificar os comentários que melhor representam uma dada série e episódio, respectivamente. Na primeira tarefa, os comentários são agrupados por série e o objetivo é classificar cada comentário à série a qual ele está associado dentre as N_s séries da nossa base de dados. Para o cálculo do TF-IDF, um documento é a coleção de comentários $\mathcal{C}^s$ associados a cada série s. Na segunda tarefa, os comentários de uma dada série são agrupados por episódio e o objetivo é classificar cada comentário ao episódio ao qual ele está associado dentre os N_e episódios dessa série. Para o cálculo do TF-IDF, um documento é a coleção de comentários $\mathcal{C}^e$ associados a cada episódio e.

Para realizar tais tarefas, utilizamos o classificador Naive Bayes[5] [Friedman et al. 1997] da coleção de ferramentas Weka [6] [Hall et al. 2009]. Foram utilizados os valores de parâmetros padrões para o Naive Bayes, e foi feita validação cruzada com 10 subconjuntos. Assim, o nosso objetivo é encontrar valores de K e α que apresentam um bom compromisso entre a acurácia da classificação e a quantidade de comentários corretamente classificados em cada uma das tarefas. Note que, se o valor de K for muito grande, muitas palavras serão usadas na tarefa de classificação e provavelmente muitas delas serão pouco discriminativas. Ao mesmo tempo, se o valor de α for muito grande, poucos comentários serão considerados na tarefa de classificação.

Os resultados dessa classificação podem ser observados na Figura 3, que mostra uma grande facilidade em se identificar a qual série cada comentário se refere. No entanto, já é possível perceber que considerar os comentários com nenhuma das palavras dentre as *top-*K presentes na representação causa uma perda considerável de acurácia. Isso segue o comportamento esperado, já que comentários desse tipo não tem um mínimo informação

[5]Como a classificação de comentários por episódio ou série não é o objetivo deste trabalho, o método selecionado não é de grande importância, tendo sido escolhido por conveniência.

[6]Disponível em `http://www.cs.waikato.ac.nz/ml/weka/index.html`



(a) Acuracia (b) Tamanho

Figura 4. A Figura 4(a) mostra a média das porcentagens de comentários corretamente classificados em seus respectivos episódios, de acordo com valores de K e α. A Figura 4(b) mostra a média das contagens de comentários considerados na classificação por episódio, de acordo com valores de K e α.

discriminativa em suas representações que permita uma classificação.

Quando passamos a classificar os comentários por episódio, dada uma série, a acurácia da classificação passa a ser menor. Esse resultado pode ser observado na Figura 4(a). Nesse gráfico, podemos observar que, conforme consideramos mais palavras (maior valor de K), torna-se mais difícil identificar a qual episódio um comentário se refere. Além disso, desconsiderar comentários com baixo valor informativo (com relação à quantidade de termos relevantes) tem um impacto positivo significativo nos resultados, principalmente para valores menores de K, embora remover muitos comentários também piore a taxa de acerto.

Por outro lado, na Figura 4(b), podemos verificar que o número de comentários considerados na avaliação cresce com valores maiores de K, e cai com valores maiores de α. Isso segue o comportamento esperado, e mostra que podemos obter um conjunto mais conciso de comentários e com maior capacidade explicativa para o episódio se escolhermos a configuração correta de parâmetros.

6.3. Explicação do resumo

Por fim, a partir dos comentários corretamente classificados, podemos escolher um subconjunto dos mesmos para explicar (ou representar) o resumo da série. Para avaliar essa tarefa, foi utilizada a métrica inspirada na ROUGE-N descrita na Seção 5.

Com base nos resultados obtidos nas tarefas de classificação, tomou-se $K = 10$ e $\alpha = 3$ como valores para os parâmetros de seleção do conjunto inicial de comentários a ser considerado, dado que foram estes os valores que retornam a maior quantidade de comentários corretamente classificados no geral. Depois, cada comentário c é representado por um conjunto de palavras $\mathcal{W}^c$ que contém as palavras presentes no comentário c e também entre as *top-*K palavras da dada tarefa de classificação. Com isso, o objetivo é encontrar o menor número possível de comentários de forma que a união de seus respectivos conjuntos $\mathcal{W}$ correspondam à totalidade do conjunto das *top-*K palavras. Esse é um caso particular do problema de *Minimum Set Cover* [Chvatal 1979], que é resolvido com bons resultados por meio de um algoritmo guloso simples, escolhendo-se sempre o comentário com maior número de palavras ainda não presentes na cobertura do conjunto.

Tabela 2. A tabela mostra para cada série: quantidade média de comentários por episódio; quantidade média de episódios selecionados para resumo por episódio; porcentagem do resumo da série R^s coberto pelo total de comentários $\mathcal{C}^s$; porcentagem do resumo R^s coberto pelos comentários selecionados $\mathcal{C}^x$.

ID Série	Comentários por episódio	Comentários selecionados por episódio	R^s coberto por $\mathcal{C}^s$	R^s coberto por $\mathcal{C}^x$
1	102,88	8,58 (8,34%)	66,42%	44,33%
19	74,02	10,06 (13,59%)	78,00%	67,14%
30	136,15	11,69 (8,59%)	78,21%	61,31%
205	69,69	6,81 (9,77%)	68,34%	50,66%
226	112,23	13,15 (11,72%)	75,00%	61,75%
356	86,04	7,54 (8,77%)	78,52%	59,74%
457	108,96	8,62 (7,91%)	84,52%	75,00%
777	166,30	9,30 (5,59%)	67,24%	51,19%
790	88,57	6,52 (7,36%)	75,66%	59,04%
820	72,60	7,92 (10,91%)	80,73%	64,25%
877	61,83	10,23 (16,55%)	79,62%	69,62%
934	135,73	8,73 (6,43%)	74,58%	54,09%
13599	165,59	9,00 (5,44%)	78,56%	59,77%

Todo esse processo é repetido para cada episódio da série a fim de escolher o menor conjunto de comentários que cobrem todas as palavras das *top-*K do episódio em questão. Por fim, tomamos a união desses conjuntos de comentários selecionados pelo *Minimum Set Cover* como os comentários selecionados como descritivos para a série.

A Figura 5 mostra o resumo retirado da Wikipédia para um dos episódios da série $s457$, tomada como exemplo ilustrativo, com as palavras marcadas de acordo com sua cobertura pelos termos encontrados nos comentários corretamente classificados ou não (parâmetros $K = 10$ e $\alpha = 3$), em verde (negrito) e vermelho (itálico), respectivamente. Tomando os resumos de todos os episódios da série em conjunto como o resumo da série em si, temos que 75% das palavras desse resumo estão presentes nos 234 comentários selecionados. Comparando-se esse resultado com a cobertura de 84% do resumo ao se considerar todos os 2833 comentários da série, podemos concluir que o método consegue, de fato, selecionar um pequeno conjunto de comentários que descrevam bem a série ($0,89$ da cobertura máxima possível pelos comentários). A Tabela 2 mostra esses mesmos resultados para as demais séries consideradas no trabalho.

7. Conclusões e Trabalhos Futuros

Este artigo apresentou uma caracterização do potencial de comentários como fonte de dados única para geração de representações formais de entidades a serem usadas por algoritmos de aprendizagem de máquina. Foram analisadas características de discussões online de episódios de séries de televisão, tais como o tamanho dos comentários, relevância das palavras, e a capacidade dos comentários para explicar um resumo extraído da

When **Ginko** was a **child**, he **lived** for a while with the *Watari*, a **group** of *nomads*, which **every year visited** a **mysterious mountain**. **Among** the *nomads*, there is **another boy**, **Isaza**, who *befriends* the **son** of the **family** who *owns* the **mountain**, and their **friendship** *grows* **despite** them **barely seeing** each other.

Figura 5. Resumo da Wikipédia para um episodio, com *stopwords* em cinza, palavras cobertas pelos comentários da série em verde e negrito, e palavras não cobertas em vermelho e itálico.

Wikipedia. Com base nessa caracterização, desenvolveu-se um método para selecionar comentários de alto valor descritivo para os episódios. O método proposto foi capaz de selecionar um pequeno conjunto de comentários que, sozinhos, conseguem descrever seus episódios e, quando tomados em conjunto, a série como um todo. Por fim, investigamos e quantificamos as diferenças entre a linguagem usada em resumos formais e manualmente gerados da linguagem usada por usuários em comentários. Como trabalho futuro, pretendemos realizar um estudo mais abrangente da relação entre o parâmetro α e o efeito de um menor corpo de comentários com o valor de cobertura do resumo. Além disso, pretendemos ampliar os atributos para a seleção e classificação de comentários considerando os métodos propostos na literatura de sumarização e ranqueamento de comentários.

Referências

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*

Bickel, S. and Scheffer, T. (2004). Multi-View Clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial Behavior in Online Discussion Communities.

Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G., and Kwon, T. T. (2015). Characterizing Conversation Patterns in Reddit. In *Proceedings of the 2015 ACM on Conference on Online Social Networks - COSN '15*, New York, New York, USA.

Chvatal, V. (1979). A greedy heuristic for the set-covering problem. *Math. Oper. Res.*, 4(3):233–235.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Mach. Learn.*, 29(2-3):131–163.

Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis : A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, number August in COLING '10, pages 340–348, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ganesan, K., Zhai, C., and Viegas, E. (2012). Micropinion generation. In *Proceedings of the 21st international conference on World Wide Web - WWW '12*, page 869, New York, New York, USA. ACM Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking comments on the social web. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 90–97. IEEE.

Ji, Y. and Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, MD.

Khabiri, E., Caverlee, J., and Hsu, C.-F. (2011). Summarizing User-Contributed Comments. In *ICWSM*.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA.

Kågebäck, M., Mogren, O., Tahmasebi, N., and Dubhashi, D. (2014). Extractive Summarization using Continuous Vector Space Models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 31–39, Gothenburg, Sweden. Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc.

Liu, C.-Y., Chen, M.-S., and Tseng, C.-Y. (2015a). IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2986–3000.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N. M., and Smith, N. A. (2015b). Toward Abstractive Summarization Using Semantic Representations. In Mihalcea, R., Chai, J. Y., and Sarkar, A., editors, *HLT-NAACL*, pages 1077–1086. The Association for Computational Linguistics.

Moratanch, N. and Chitrakala, S. (2016). A survey on abstractive text summarization. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–7. IEEE.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14.

Potthast, M. and Becker, S. (2010). Opinion Summarization of Web Comments. pages 668–669.

Siersdorfer, S., Chelaru, S., Pedro, J. S., Altingovde, I. S., and Nejdl, W. (2014). Analyzing and Mining Comments and Comment Ratings on the Social Web. *ACM Trans. Web*.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

von Ahn, L. and Dabbish, L. (2004). Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04.

Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. Y. (2015). Network Representation Learning with Rich Text Information. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15.

Yang, Z., Cai, K., Tang, J., Zhang, L., Su, Z., and Li, J. (2011). Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 255, New York, New York, USA. ACM Press.

yew Lin, C. (2004). Rouge: a package for automatic evaluation of summaries. pages 25–26.