A referenciação em textos do YouTube: um estudo com vistas à análise de sentimentos

Referenciation in YouTube Texts: a study for Sentiment Analysis

Alexandre Ribeiro Afonso*

RESUMO

Este artigo descreve um estudo sobre o fenômeno da referenciação a partir de um *corpus* textual extraído do YouTube. O objetivo é verificar a utilização dos referentes em postagens positivas, negativas ou neutras, e como essa verificação contribuiria para as técnicas de análise de sentimentos automática ou semiautomática. Emprega-se a análise de conteúdo e do discurso como métodos de pesquisa, além de *software* de análise estatística para coleta de frequências de palavras. Conclui-se que a identificação de aspectos de comunicação e informação, tais como os referentes mais utilizados, a não pureza de postagens positivas e negativas, além, da necessidade de criar um critério para a interpretação de dados textuais, pode aperfeiçoar o processo de análise de sentimentos.

Palavras-chave: Análise de Sentimentos; Análise de Conteúdo; *Corpus*; Referenciação.

ABSTRACT

This article describes a study about the phenomenon of referenciation by using a textual corpus extracted from YouTube. The purpose is to verify how the referents appear in positive, negative, or neutral posts and how this verification can affect the techniques directed to automated or semi-automated sentiment analysis. We applied content and discourse analysis as research methods, and statistical analysis software for the collection of word frequencies. The conclusion is that the identification of communication and information aspects, such as the most used referents, the non-purity of positive and negative posts, and the need to create a criterion for interpreting textual data can improve sentiment analyses process.

Keywords: Sentiment Analysis; Content Analysis; Corpus; Reference.

INTRODUÇÃO

Um campo de pesquisa que vem se desenvolvendo desde o início deste século é a análise de opiniões (automatizada ou semiautomatizada),[1] a partir de dados textuais gerados em mídias sociais digitais.

Segundo Afonso e Té (2017), desde que a postagem de conteúdos se tornou tecnicamente ágil e ubíqua, há o interesse neste tipo de extração de conhecimento, sobre produtos, serviços, eventos ou entidades citados nas mídias digitais. A elaboração de técnicas para obtenção de quadros de opinião, a partir das postagens

* Doutor em Ciência da Informação pela Universidade de Brasília (UnB). E-mail: rafonso.alex@gmail.com.

[1] Análise de sentimentos e mineração de opiniões são termos comumente empregados quando a análise de opiniões é automática.

em mídias sociais digitais, é um desafio que tem sido proposto ao meio acadêmico e empresarial.

As mídias sociais como o Twitter, o Facebook e o YouTube, entre outras, são fontes consideráveis de dados textuais para a identificação de sentimentos sobre temas diversos, por meio da exploração de suas mensagens opinativas. Especialistas em análise de dados, com ferramentas específicas, levantam medidas estatísticas e constroem gráficos sobre tendências de opinião. Em outra vertente, é possível programar sistemas inteligentes que, automaticamente, identificam padrões de sentimentos no texto, tais como positividade, negatividade, medo, raiva, tristeza, entre outros. Com o objetivo de substituir um analista de dados humano, várias técnicas são propostas para tal inteligência computacional (ARAÚJO; GONÇALVES; BENEVENUTO, 2013).

Nessa atividade de análise (seja automática ou semiautomática), não há dúvidas de que a compreensão e a descrição da maneira como os internautas se expressam no texto, em uma língua específica, ao comunicarem suas opiniões, é uma etapa significativa do processo. Como exemplo dessa descrição, pode-se citar a mensagem propagada no Twitter, a qual tem característica peculiares: é curta, com menos caracteres que em outras mídias, o que faz o discurso objetivo e conciso, para garantir uma comunicação eficaz. Em outras mídias, o texto pode ser longo, e o número de referentes (entidades de quem se fala) e seus qualificadores podem ser extensos, com explanações diversas.

Um aspecto importante a ser observado é que as mensagens são codificadas em língua nativa (em português do Brasil). Essa língua, nesse ambiente das mídias sociais digitais, e para o propósito de análise de dados, ainda é estudada discretamente. Se considerarmos tanto estudos linguísticos, em informação e comunicação, ou na área tecnológica, a quantidade de pesquisas sobre o tema é baixa quando comparada a outras línguas. Os estudos sobre o uso da língua nesses ambientes de discussão podem contribuir de maneira significativa para a compreensão da estruturação da opinião, e a partir daí, pode-se construir variedades mais robustas de *software* para auxílio ou execução da análise de sentimentos.

Este artigo descreve, especificamente, os comentários no YouTube. A maioria dos estudos da mensagem em redes sociais para o português brasileiro ocorre para o Facebook e o Twitter. No YouTube, diferentemente de outras mídias, há referências ao vídeo que acompanha os comentários, ao contrário do Twitter e do Facebook, que não possuem, necessariamente, um vídeo específico como guia das discussões. A pesquisa procurou descrever os referentes, e suas características em uso, em mensagens positivas, negativas e neutras ocorrentes nos comentários de um vídeo específico: uma crítica ao filme *Batman versus Superman*: *a origem da justiça*. Ele foi exibido nos cinemas em 2016 no Brasil, e gerou uma polêmica abrangente entre os fãs, com a polarização de opiniões nas mídias e na crítica especializada.

Por meio deste estudo, é possível verificar alguns padrões sobre o fenômeno da referenciação nas postagens ligadas ao vídeo. Tais descrições contribuem para a detecção das dificuldades encontradas por métodos automáticos ou semiautomáticos, ao identificarem a polaridade e sentimentos contidos no texto opinativo.

PROBLEMÁTICA E OBJETIVOS DA PESQUISA

As áreas, subáreas do conhecimento e especialidades que têm o texto como fonte de dados para pesquisa são diversas. Seja nas humanidades, ciências exatas ou ciências da saúde, verifica-se que, por diferentes métodos, o texto pode ser analisado para se chegar a uma consideração científica, tendo um *corpus* como fonte de dados, com a finalidade de responder às questões de uma investigação.

Alguns campos científicos, como as ciências sociais, podem estar interessados em captar e analisar o conteúdo das mensagens, com o objetivo de compreender o seu significado, por vezes, dentro de um contexto social. Ou ainda, o objetivo pode ser compreender além do significado da mensagem: o pensar e a intenção comunicativa do emissor ou o impacto do conteúdo no receptor. Nesse caso, o formato do texto que carrega a mensagem pode ser o de respostas a um questionário, um texto jornalístico, entre outros formatos. Segundo Moraes (1999), a análise de conteúdo é uma ferramenta, um guia prático para a ação, sempre renovada em função dos problemas cada vez mais diversificados que se propõe a investigar. Pode-se considerá-la como um único instrumento, mas marcado por uma grande variedade de formas e adaptável a um campo de aplicação muito vasto: a comunicação.

A ciência linguística relata estudos que podem ter interesse não somente no conteúdo da mensagem, mas no funcionamento do próprio código linguístico. O texto escrito ou falado, nesse caso, pode ser estudado para compreender o uso da língua no processo comunicativo. Sob a ciência linguística, a forma, e não somente o conteúdo, será objeto de pesquisa. Para Oliveira (2009), a linguística de *corpus* pode ser considerada como "a face moderna da linguística empírica", sendo a linguagem vista como um fenômeno social e analisada a partir de atos concretos de comunicação, isto é, textos reais, buscando o significado onde este é negociado, ou seja, no discurso.

A informática e a ciência da informação também trabalham com o texto, na busca de inovações tecnológicas. A informática estaria focada na construção de tipos de *software* para identificar sentimentos em textos, entre outras aplicações que manipulam dados textuais (ARAÚJO; GONÇALVES; BENEVENUTO, 2013). Já a ciência da informação, em sua atuação clássica, procuraria compreender o processo de transformação de dados textuais em conhecimento. Segundo Schiessl (2007), tem-se chamado essa atividade investigativa de descoberta de conhecimento em textos. Ela objetiva automatizar o processo de transformar dados textuais em informação para possibilitar a aquisição de conhecimento.

Ainda como método de pesquisa, considerando o texto como fonte de análise, para Caregnato e Mutti (2006), a análise do discurso pode ser considerada um método que visa a descoberta do sentido do texto. O analista, ao utilizar tal análise, fará uma leitura do texto enfocando a posição discursiva do sujeito, legitimada socialmente pela união do social, da história e da ideologia, produzindo sentidos. Enquanto a análise do discurso busca os efeitos de sentido relacionados ao discurso, a análise de conteúdo fixa-se apenas no conteúdo do texto, sem fazer relações além deste.

Este estudo acaba por abranger três campos: linguística, informática e comunicação e informação. Na análise de sentimentos citada, pouco se considerou o texto como tópico central de pesquisa, mas antes de construir um aplicativo inteligente para a detecção de positividade ou negatividade contida no texto opinativo, a caracterização dessa informação deve ser investigada, suas peculiaridades, tendências de uso da língua e casos atípicos. Nesse viés, descreve-se aqui o fenômeno da referenciação nas postagens. O referente é uma entidade presente no

texto, ou seja, o objeto de quem se fala. No caso do filme que é alvo de críticas, a procura visa captar os referentes em comentários (o filme, os personagens, as cenas, etc.).

Na linguística textual, o fenômeno da referenciação pode ser compreendido segundo o trabalho de Koch (2008). A autora descreve que o primeiro passo na construção de um texto é a introdução de um objeto de discurso na memória textual (em geral, por meio de um nome próprio ou forma nominal). Isto é, um novo objeto de discurso é construído e introjetado na memória, onde vai preencher um nódulo, ou seja, passar a ter um endereço cognitivo, de modo a ficar em foco e disponível para retomadas ou remissões. Quando a introdução se faz por meio de um nome próprio, tem-se apenas a nomeação do objeto. Já no caso de se tratar de uma expressão nominal, opera-se uma primeira categorização do objeto de discurso, o qual, a cada retomada, pode ser mantido como tal ou, então, recategorizado por outras expressões nominais.

Deve-se considerar, portanto, que pode haver uma renomeação constante do referente nas postagens. O referente "Filme *Batman versus Superman*", por exemplo, pode estar colocado com várias outras nomeações: "Filme", "Filminho Ruim", "Filmaço", "Ele", "Uma Boa Ideia", etc. Observa-se ainda, nos exemplos dados, que ao renomear, a opinião já pode estar aglutinada ao nome ou expressão nominal utilizada.

Nas mídias sociais, como o YouTube, os referentes representados como objetos do discurso, modificam-se nas diversas postagens, e não somente no texto de uma única postagem. O trabalho que, nessa direção, mais se aproxima desta pesquisa é o recente estudo de Afonso e Té (2017), os quais descrevem justamente as formas nominais para o "processo de *impeachment*" da ex-presidente Dilma Rousseff, iniciado em 2015. As postagens opinativas tanto positivas quanto negativas são descritas, considerando a maneira como esse referente "processo de *impeachment*" se camufla em formas nominais diversas no *corpus* coletado. Este estudo capta os resultados de modificação de um único referente "processo de *impeachment*" nas postagens do YouTube em três vídeos.

No trabalho aqui descrito, procura-se justamente identificar quais referentes aparecem nas postagens, e como é a distribuição desses referentes nos vários textos dos usuários.

MATERIAIS E QUESTÕES DE PESQUISA

Materiais de análise

A compreensão do fenômeno da referenciação ocorre nas postagens para um vídeo específico do YouTube.[2] Este vídeo contém uma série de críticas sobre o filme *Batman versus Superman*: *a origem da justiça*. O vídeo contém a fala de três críticos que fazem observações sobre o filme: ora positivas, ora negativas. A polêmica criada sobre o produto, construída inicialmente pela crítica americana, torna a discussão produtiva, com apontamentos de aspectos variados.

Os três participantes constroem um ambiente descontraído, com um linguajar pouco formal, utilizando expressões e terminologias do universo *nerd*/*geek*, dos filmes

[2] Disponível em:<www.youtube.com/watch?v=FrsKsV2aSyE>.

contemporâneos de super-heróis e HQs, tais como: *Fan service*, *Crossover*, *CGI*, entre outros termos e expressões características desse domínio. Os críticos demonstram conhecimentos sobre a história dos personagens e produções cinematográficas do gênero, ao apontarem coerências e incoerências, e sobre os universos das produtoras de filmes baseados em HQs: as companhias Marvel e DC Entertainment, criadoras dos principais filmes e HQs de super-heróis atuais. O uso de palavrões é constante, para enfatizar tanto aspectos positivos quanto negativos do filme. O vídeo da crítica possui 11 minutos e 46 segundos de duração.

O *corpus* de análise foi construído retirando-se, aleatoriamente, uma amostra de 412 comentários do vídeo citado – o total observado era de 6.110 postagens de comentários. Evitaram-se postagens que respondiam outras postagens, pois estas podem suportar conteúdos que fogem da análise fílmica e envolvem a criação processos de conversação, muitas vezes criando novos temas de discussão. As postagens que envolviam conteúdo fora do domínio considerado (como propagandas totalmente fora do universo fílmico, ou do vídeo da crítica) foram substituídas. O *corpus* final, após as filtragens e a retirada de postagens repetidas, armazenou 409 postagens, sendo 182 positivas, 184 negativas e 43 neutras.

Questões de pesquisa

As seguintes questões de pesquisa foram levantadas:

a) Quais referentes são opinados no *corpus* coletado? Eles podem ser colocados em categorias?

b) Como estabelecer um critério de positividade, negatividade e neutralidade para as postagens, já que uma única postagem pode conter diversos referentes?

c) A partir do critério estabelecido na questão 2, quantificar os referentes opinados para as postagens positivas, negativas e neutras. Há similaridade nos valores das quantificações?

d) A partir do critério estabelecido na questão 2, quantificar as postagens positivas que possuem opiniões negativas, e quantificar as postagens negativas que possuem opiniões positivas.

e) Existem subjetividades no julgamento de positividade e negatividade sobre os referentes? Qual a natureza de tais subjetividades?

MÉTODOS DE PESQUISA E RESULTADOS

O método de pesquisa da Análise de Conteúdo descrito em Moraes (1999) e Bardin (1977) foi empregado, uma vez que o objetivo é a criação de categorias para compreensão das mensagens utilizadas, mas há também uma análise interpretativa dos textos para a identificação de algumas características. A análise de conteúdo foi auxiliada com o uso do *software* AntConc para análise de *corpus,* descrito por Kader e Richter (2013), e o Microsoft Excel. A seguir, para cada questão de pesquisa, descrevem-se as etapas do método adotado.

a) Quais referentes são opinados no *corpus* coletado? Eles podem ser colocados em categorias?

A partir das 409 postagens, os referentes foram classificados em categorias, observa-se que uma postagem pode ter mais de um referente, mas um referente pode estar somente em uma categoria. As categorias foram criadas baseando-se na leitura e releitura do material até que elas fossem visíveis. Um referente é identificado num texto postado quando uma opinião sobre ele surge. Nota-se que uma postagem pode conter vários referentes e opiniões diversas sobre eles, independentemente de a opinião ser positiva, negativa ou neutra. O Quadro 1, a seguir, exibe as categorias de referentes encontradas.

Quadro I – Descrição das categorias de referentes encontradas no *corpus*.

Nome da Categoria	**Exemplo**
Ator de outro filme específico	[...] Christopher Nolan e **Christian Bale** imortalizaram o Batman!
Ator específico	**Ben Affleck** não me convenceu !!
Cena específica	[...] **aquela cena da liga da justiça** foi sem noção [...]
Cena (em sentido genérico)	Ví furos de roteiros, cortes bruscos de **cena editada**, atitudes incoerentes [...]
Cenas específicas	**As cenas com ela** são mito.
Cenas (em sentido genérico)	[...] exageros para todos os lados, **cenas confusas** [...]
Cinematografia do filme	Mano esse filme eu esperava mais [...] A **cinematografia** é excelente.
Crítica específica	Na moral que não consigo levar a sério **uma crítica que defenda esse Lex maluco** [...]
Crítica (em sentido genérico)	Filme fraco. **Crítica** sem critério.
Críticas específicas	Ótimo vídeo! Boas **pontuações**.
Críticas (em sentido genérico)	Realmente não estou entendo pq as **críticas** estão pesando tanto esse filme!
Crítico específico	Concordei **com o do meio**, aquela cena da liga da justiça foi sem noção [...]
Críticos específicos	Não estou criticando o vídeo de vcs, blz? Sou fã demais **de vcs**! Bjs!
Críticos (em sentido genérico)	Acho q esses **críticos de meia tigela** tinham q se interar mais sobre o universo DC.
Direção de outro filme específico	**Christopher Nolan** e Christian Bale imortalizaram o Batman!
Direção do filme	Parem de chamar o **Snyder** para dirigir os filmes de heróis POR FAVOR [...]
Edição do filme	O filme é excelente, bem dirigido, bem

Nome da Categoria	Exemplo
	editado e bem contado.
Efeito (em sentido genérico)	**Muita câmera lenta e tela preta**, forçava uma tensão que não estava funcionando.
Efeitos especiais	**Computação gráfica** ficou a desejar.
Elemento do personagem	[...] **aquele batmóvel** foi retardado [...]
Fan service	[...] colocando o **fan service** todo mundo sai feliz, eu incluída.
Filme BvS	Cara, não gostei do **filme** [...]
Grupo da plateia	[...] eu vi a piração **dos nerds** ontem na sala do cinema !!!!
HQ	[...] eles claramente se inspiraram em uma **HQ** BOA.
Montagem do filme	Fora a **montagem do filme**, que teve cortes demais, não dava tempo de assimilar [...]
Outro filme específico	Tenho gostado cada vez menos dos filmes da Marvel, especificamente depois do **Homem de Ferro 3**
Outros desenhos	Eu não sou fãnzona da Marvel ou desses filmes ou **desenhos** do tipo.
Outros filmes	[...] gostei muito do filme, tô ansioso para ver os **próximos**[...]
Parte do filme	[...] O filme só começou a funcionar na **segunda metade**[...]
Parte do roteiro	[...] só achei estranho **esse negocio deles brigarem** [...]
Parte dos críticos	**Esses críticos americanos** são um bando de [...]
Partes do filme	[...] e **outros pontos** não me agradou .
Partes do roteiro	Tem **diálogos** confusos que precisam de interpretação [...]
Personagem específico	Achei que o **Lex** interpretou o coringa [...]
Personagem específico de outro filme	Cara eu quero ver o **capitão** [...] ansioso demais.
Personagens específicos	Filme fraco, cheio de furos. **Heróis** acéfalos.
Personagens (em sentido genérico)	[...] e olha que eu amo os **personagens**.
Plateia (em sentido genérico)	**As pessoas** estão perdendo a habilidade de se divertir.

Nome da Categoria	Exemplo
Produtora específica	A **DC** entregou um ótimo filme.
Roteirista do filme	Vc vê claramente que o diretor e **roteirista** se perdem [...]
Roteiristas do filme	**Eles** sempre favorecem o Batman, na luta entre eles toda vez.
Roteiro (em sentido genérico)	Achei a **história** esquisita, mas as lutas foram legais.
Tempo do filme	[...] me pareceu que o filme precisava de **mais tempo** para apresentar as coisas.
Trailer do filme	O filme eh ruim muito ruim e o **trailer** entregou tudo.
Trailers (em sentido genérico)	[...] Não fui no cinema por que ja não gostei dos **trailers**.
Trilha sonora	A **trilha sonora** foi um lixo [...]
Universo dos super-heróis	[...] **universo** totalmente surreal o qual alimentava minha imaginação quando criança.
Vídeo da crítica	Mano, faz exatamente um ano que e aasisti essa mitagem **de vídeo**, que massa.

Fonte: elaborado pelo autor, adaptado de: www.youtube.com/watch?v=FrsKsV2aSyE.

b) Como estabelecer um critério de positividade, negatividade e neutralidade para as postagens, já que uma única postagem pode conter diversos referentes?

Para responder a tal questão, o material do *corpus* passou por uma análise discursiva e verificou-se que existem muitas maneiras de se estabelecer positividade, negatividade e neutralidade em uma postagem. Uma possibilidade é considerar não a polaridade da postagem como um todo, mas considerar as opiniões por referente em todo o *corpus* de pesquisa, e contar, a partir daí, a polaridade da opinião sobre os referentes identificados.

Há uma diferença considerável nesses dois tipos de análise: a contagem de polaridades por referente revela quais e quantos referentes são opinados como positivo, negativo ou neutro no *corpus*. Já uma contagem por postagens revela quais e quantos usuários opinaram como positivo, negativo ou neutro, e a análise de dados (automática ou semiautomática), neste último caso, efetua-se por postagem. Se a postagem for considerada como elemento de avaliação da polaridade, leva-se em consideração que uma postagem pode ter múltiplos referentes e carrega positividades e negatividades associadas a tais referentes. Ou ainda, para um mesmo referente podem existir aspectos positivos e negativos.

Além disso, observa-se que alguns referentes indicam superioridade na marca da opinião, que se sobrepõe a outras opiniões de outros referentes. Por exemplo, quando um usuário comenta que "amou demais o filme" e na mesma postagem comenta que "não gostou do carro do Batman", uma possibilidade plausível de

interpretação é que o usuário considerou positivo o conjunto da obra, mas destacou apenas um item negativo. Logo, a postagem é de opinião positiva, mas com restrição. Porém, outro analista de dados poderia considerar que há um empate de positividade e negatividade na postagem, com uma opinião positiva e uma negativa para a mesma postagem, não havendo, portanto, uma definição clara de *polaridade* para a postagem do usuário na forma integral – ela depende da percepção do analista.

Há ainda outras questões a serem tratadas. A neutralidade na postagem ocorre de que maneira? Um mesmo número de opiniões positivas e negativas numa postagem leva a uma interpretação de neutralidade? Ou depende do referente, do fato de ele ser de uma hierarquia superior? Uma outra questão, ainda, portanto, pode ser levantada: o "vídeo da crítica" no YouTube também seria uma entidade de nível hierárquico maior, por conter elementos próprios que são opinados, tais como: críticas do vídeo e críticos do vídeo?

Nessa perspectiva, o que se observa, de acordo com a interpretação do material, é que um critério de julgamento deve ser estabelecido antes de qualquer tipo de análise de dados. O critério utilizado vai depender muito do que o analista de dados almeja descobrir: ele anseia descobrir a opinião dos usuários? Ou ele quer conhecer sobre os referentes citados, independentemente de quem opina? Sobre quais referentes ele quer descobrir? O peso de opinião é o mesmo para todos os referentes, para se definir a polaridade de uma postagem?

Considerando tais questões, e considerando o referente "Filme *Batman versus Superman*" como o elemento principal de análise para esta pesquisa, é visto que o critério de escolha da polaridade para uma postagem se baseia em uma hierarquia, em que a positividade, negatividade ou neutralidade é indicada primeiramente pela opinião em relação ao referente "Filme *Batman versus Superman*". Isso quer dizer que se este referente for julgado positivamente na postagem, independentemente de existirem outros referentes com opinião associada, a postagem será considerada de polaridade positiva. Da mesma maneira, se o "Filme *Batman versus Superman*" for visto negativamente pelo usuário, a postagem será, pela hierarquia adotada, de polaridade negativa.

Sob essas observações, pode-se estabelecer o seguinte critério para o material coletado e para o julgamento da polaridade de uma postagem, critério que pode ser descrito na forma de regras, da seguinte maneira:

Uma postagem é classificada como "positiva":

- Se existirem somente positividades para os referentes no texto postado.
- Se existir a positividade do referente de hierarquia maior ou dominante (no caso, foi considerado dominante "Filme *Batman versus Superman*") com aspectos negativos ou neutros para referentes de menor hierarquia em relação ao referente dominante.
- Se só existirem referentes de menor hierarquia que "Filme *Batman versus Superman*", com maior número de positividades para esses referentes.

Uma postagem é classificada como "negativa":

- Se existirem somente negatividades para os referentes no texto postado.
- Se existir a negatividade do referente de hierarquia maior ou dominante (no caso, foi considerado dominante "Filme *Batman versus Superman*") com aspectos positivos ou neutros para referentes de menor hierarquia em relação ao referente dominante.

- Se só existirem referentes de menor hierarquia que "Filme *Batman versus Superman*", com maior número de negatividades para esses referentes.

Uma postagem é classificada como "neutra":

- Se há somente referentes de mesmo nível hierárquico e a subtração do número de opiniões positivas pelo número de opiniões negativas sobre os referentes resulta zero.
- Se existirem na postagem somente referentes sem sentido positivo ou negativo, seja o referente dominante ou não, isso quer dizer que há somente outro tipo de sentido para os referentes.

Exemplo:

Quadro 2 – Exemplos de polaridade para as postagens do *corpus* segundo o critério adotado.

Postagem	**Polaridade**	**Razão da Polaridade**
Daora Batman v Superman, só achei estranho esse negocio deles brigarem...	Positiva: existência do referente dominante positivado.	O primeiro referente pertence ao grupo "Filme BvS", representado na postagem por "Batman vs Superman", sendo o referente dominante e positivado com o adjetivo "*Daora*". Porém, observa-se que há uma negatividade no texto restante, associada ao referente "Esse negócio deles brigarem" que pertencente ao grupo de referentes "Parte do roteiro".Essa negatividade é colocada na oração: "*só achei estranho esse negócio deles brigarem*".
O excesso de Lois Lane ferrou o roteiro, e o desfecho da mãe foi podre	Negativa: três referentes não dominantes negativados.	O primeiro referente "*Lois Lane*" pertence ao grupo de referentes "personagem específico" e é negativado em "*O excesso de Lois Lane ferrou o roteiro*". O referente "*o roteiro*" pertence ao grupo "Roteiro genérico", sendo também negativado na oração "*O excesso de Lois Lane ferrou o roteiro*". O referente "*o desfecho da mãe*" pertence ao grupo de referentes "Parte do filme" e é negativado na oração "*o desfecho da mãe foi podre*".
O Filho de Krypton vs. O Morcego de Gotham	Neutra: não tem polaridade definida.	A postagem oferece apenas uma constatação do usuário, sem positividade ou negatividade, sobre o referente do grupo "Filme BvS". O usuário apenas renomeia o filme.

Fonte: elaborado pelo autor, adaptado de www.youtube.com/watch?v=FrsKsV2aSyE

c) A partir do critério estabelecido na questão 2, quantificar os referentes opinados para as postagens positivas, negativas e neutras. Há similaridade nos valores das quantificações?

A seguinte quantificação foi obtida:

Tabela I – Número de ocorrências para cada categoria de referentes com destaque para as de maior quantidade.

Categoria	Neg.	Pos.	Neu.
Filme BvS	**183**	**243**	**17**
Personagem específico	**73**	**72**	**20**
Roteiro genérico	**31**	11	**4**
Crítico específico	**30**	**36**	2
Outro filme específico	**26**	**17**	2
Parte do roteiro	**26**	**16**	1
Cena específica	18	6	2
Elemento do personagem	17	9	0
Produtora específica	11	12	**6**
Partes do filme	9	**16**	0
Vídeo da crítica	8	5	1
Direção do filme	8	1	0
Cenas específicas	7	9	1
Ator específico	7	5	0
Parte do filme	6	10	**4**
Crítica específica	5	0	0
Críticos específicos	5	0	0
Grupo da plateia	4	14	**5**
Efeitos especiais	4	5	0

Fonte: elaborado pelo autor.

Na tabela anterior, observa-se que os comentários são principalmente para as categorias de referentes: "FilmeBvS", "Personagem específico", "Roteiro genérico", "Crítico específico", "Outro filme específico", "Parte do roteiro". Independentemente da polaridade da opinião, a concentração das opiniões se estabeleceu nessas categorias.

d) A partir do critério estabelecido na questão 2, quantificar as postagens positivas que possuem opiniões negativas e quantificar as postagens negativas que possuem opiniões positivas.

Foi observado nas postagens positivas que de 182 postagens, 83 têm pelo menos 1 referente negativado ou, efetuando os arredondamentos decimais, 45,6% das postagens consideradas positivas contêm alguma opinião negativa para os referentes.

Foi observado também nas postagens negativas que de 184 postagens, 45 têm pelo menos um referente positivado. Isso quer dizer que, efetuando os arredondamentos decimais, 24,5% das postagens consideradas negativas possuem alguma opinião positiva para os referentes.

e) Existem subjetividades no julgamento de positividade e negatividade sobre os referentes? Qual a natureza de tais subjetividades?

Utilizando uma análise discursiva sobre o *corpus* de estudo, foi possível coletar alguns casos de interpretações difíceis de serem consideradas conclusivas:

1) *Resumo, dois viram erros e acertos. Um achou que foi o melhor filme da história e não deixava ninguém discordar.*
2) *Muito melhor q vingadores.*
3) *Eu não gostava do Superman, mas agora com essa porrada que ele levou, o achei mais humano e o entendi.*
4) *Colcha de retalhos.*
5) *Depois de ver este filme penso: o filme do Pelé foi melhor.*

No exemplo 1, observa-se uma opinião sobre um dos críticos: "*Um achou que foi o melhor filme da história e não deixava ninguém discordar*". Seria essa opinião sobre o crítico de polaridade negativa ou neutra?

No exemplo 2, o enunciado claramente coloca o filme *Batman versus Superman* acima de outro filme, mas isso quer dizer que a opinião contida é positiva? Poder-se-ia argumentar que essa descrição não forma uma opinião, já que a positividade é apenas acima de outro filme.

No exemplo 3, o usuário coloca "[...] *mas agora com essa porrada que ele levou, o achei mais humano e o entendi*". Isso quer dizer que agora ele simpatiza com o personagem Superman e sua visão sobre o personagem é positiva?

No exemplo 4, a predicação "colcha de retalhos" é uma opinião negativa ou neutra?

No exemplo 5, claramente há na mensagem um tom de sarcasmo, mas será que se pode considerar tal opinião negativa em relação ao filme?

Observa-se, portanto, que diferentes possibilidades de interpretação podem ocorrer no julgamento da polaridade, e algumas escolhas arbitrárias são feitas pelo analista de dados. O texto em linguagem natural pode não ser claro por motivos diversos: a falta de informação suficiente no texto para uma interpretação exata, ideias contraditórias, ambiguidades interpretativas, ironias e sacarmos, que não permitem uma classificação de polaridade exata.

Quando a análise do texto é executada via *software* (automática), a complexidade é maior, já que um computador teria dificuldades significativas para compreender metáforas, ironias e sarcasmos. A quantidade de conhecimento da máquina teria que ser muito grande e representada de forma eficaz para a compreensão desse tipo de mensagem. Tal compreensão depende de conhecimentos culturais, como no exemplo 5, para julgar a polaridade é preciso saber quem é, e o que o representa para

o Brasil o nome "Pelé". Além disso, é necessário o conhecimento que esse bordão se tornou um "meme" nas redes sociais.

CONSIDERAÇÕES FINAIS

A pesquisa relatada neste artigo teve como objetivo a descrição do fenômeno da referenciação nas postagens do YouTube, para um vídeo específico. A descrição de estratégias de comunicação envolvendo os referentes identificados levaria a métodos de análise de sentimentos com melhores resultados.

Em uma primeira visualização dos dados textuais, é observada a persistência de uma linguagem característica para esse tipo de vídeo, tanto no vocabulário falado dos críticos no vídeo quanto dos usuários nas postagens. O linguajar da "cultura *nerd/geek*" e as expressões informais são constantes. Devido a isso, coloca-se que o levantamento do vocabulário utilizado nesse ambiente, para indicar positividade e negatividade (ou ainda outros sentimentos), deve ser uma das etapas iniciais para a construção de um método eficaz de análise de dados.

De acordo com as respostas obtidas para as questões 1 e 3 da pesquisa, verifica-se que um grupo de referentes é recorrente nas postagens. Podem camuflar-se em várias formas quando são colocados nas postagens, mas o conjunto é limitado, independentemente da polaridade da opinião nas postagens.

A resposta à questão de pesquisa 2 mostra que compreender a informação registrada antes da criação de qualquer método, automático ou semiautomático, de análise de dados pode levar a métodos mais elaborados de acordo com o conteúdo dos registros. Antes de qualquer análise de dados, é preciso especificar quais referentes interessam ao levantamento de opinião pública (o vídeo do YouTube? Os críticos? O filme? Etc.). Em seguida, é necessário elaborar um critério sobre que é positividade, negatividade e neutralidade para o ambiente estudado, e quais elementos (referentes) sob alvo de opiniões são mais relevantes para a análise.

Observa-se, ainda, nos resultados descritos pela resposta da questão 4, que postagens positivas e negativas não são puras. Isso quer dizer que as postagens positivas frequentemente contêm referentes negativados em seu corpo, assim como postagens negativas frequentemente contêm referentes positivados.

Na reposta da última questão, de número 5, verifica-se que a vagueza na informação, a ambiguidade, o sentido metafórico, sarcástico e irônico ocorrem nas postagens, o que pode levar a uma subjetividade e dificuldade de interpretação do texto (algo característico de enunciados não formais em linguagem natural).

O estudo da informação e dos processos de comunicação levaria a descrições de uso da língua e expressões características em ambientes específicos. Como observado, pode-se, dessa maneira, construir métodos de análise de dados mais eficientes e eficazes para os ambientes midiáticos digitais de interação.

Artigo recebido em 16/07/2017 e aprovado em 26/10/2017.

REFERÊNCIAS

AFONSO, Alexandre Ribeiro; TÉ, Jordão. Um estudo sobre referenciação e a construção da opinião a partir de um corpus textual extraído do YouTube. *Domínios de Lingu@gem*, v. 11, n. 2, p. 339-350, 2017.

ARAÚJO, Mateus; GONÇALVES, Pollyanna; BENEVENUTO, Fabrício. Métodos para análise de sentimentos no Twitter. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, WEBMEDIA 2013, 19., Salvador, 2013. *Proceedings*... Salvador: Brazilian Computer Society, 2013.

BARDIN, Laurence. *Análise de conteúdo*. Lisboa: Edições 70, 1977.

CAREGNATO, Rita Catalina Aquino; MUTTI, Regina. Pesquisa qualitativa: análise de discurso *versus* análise de conteúdo. *Texto Contexto Enfermagem*, v. 15, n. 4, p. 679-84, 2006.

KADER, C. C. C.; RICHTER, M. G. Linguística de *corpus*: possibilidades e avanços. *Instrumento*: revista de estudo e pesquisa em educação, v. 15, n. 1, p. 13-23, jan./jun. 2013.

KOCH, I. V. Como se constroem e se reconstroem os objetos-de-discurso. *Investigações*, Recife, v. 21, n.2, p. 99-114, 2008.

LAURENCE, Bardin. *Análise de conteúdo*. Lisboa: Edições 70, 2004.

MORAES, Roque. Análise de conteúdo. *Revista Educação*, Porto Alegre, v. 22, n. 37, p. 7-32, 1999.

OLIVEIRA, Lúcia P. Linguística de *corpus*: teoria, interfaces e aplicações. *Matraga*, v. 16, n. 24, 2009.

RECUERO, Raquel. Discutindo análise de conteúdo como método: o # DiadaConsciênciaNegra no Twitter. *Cadernos de Estudos Linguísticos*, v. 56, n. 2, 2014.

SCHIESSL, J. M. *Descoberta de conhecimento em texto aplicada a um sistema de atendimento ao consumidor*. Brasília, 2007. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação, Departamento de Ciência da Informação e Documentação, Universidade de Brasília.