UNIVERSIDADE SAGRADO CORAÇÃO

VINICIUS RIOS GIANEZI

MINERAÇÃO DE OPINIÃO APLICADA PARA A CLASSIFICAÇÃO DE CRÍTICAS EM PORTUGUÊS SOBRE CINEMA

BAURU
2014

VINICIUS RIOS GIANEZI

MINERAÇÃO DE OPINIÃO APLICADA PARA A CLASSIFICAÇÃO DE CRÍTICAS EM PORTUGUÊS SOBRE CINEMA

Trabalho de conclusão de curso apresentado ao Centro de Ciências Exatas e Sociais Aplicadas como parte dos requisitos para obtenção do título de bacharel em Ciência da Computação, sob a orientação do Prof. Me. Patrick Pedreira Silva.

BAURU
2014

Gianezi, Vinicius Rios.

G433m

Mineração de opinião aplicada para a classificação de críticas em português sobre cinema / Vinicius Rios Gianezi. -- 2014.

58f. : il.

Orientador: Prof. Me. Patrick Pedreira Silva.

Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade do Sagrado Coração – Bauru – SP.

1. Mineração de opinião. 2. Classificação de críticas. 3. Análise de sentimento. I. Silva, Patrick Pedreira. II. Título.

VINICIUS RIOS GIANEZI

MINERAÇÃO DE OPINIÃO APLICADA PARA A CLASSIFICAÇÃO DE CRÍTICAS EM PORTUGUÊS SOBRE CINEMA

Trabalho de conclusão de curso apresentado ao Centro de Ciências Exatas e Sociais Aplicadas como parte dos requisitos para obtenção do título de bacharel em Ciência da Computação, sob a orientação da Prof. Me. Patrick Pedreira Silva.

Banca examinadora:

_____
Prof. Me. Patrick Pedreira Silva
Universidade do Sagrado Coração

_____
Prof. Dr. Elvio Gilberto da Silva
Universidade do Sagrado Coração

_____
Prof. Esp. André Luiz Ferraz Castro
Universidade do Sagrado Coração

Bauru, 29 de novembro de 2014.

RESUMO

No atual cenário mundial, marcado pela globalização e pelo aumento constante da competitividade nas empresas, a grandiosa massa de dados gerada diariamente na web comporta uma rica fonte de opiniões de usuários acerca dos mais variados assuntos, que inclui avaliações de produtos e serviços disponíveis no mercado. Neste contexto, surge a necessidade de automatização do processo de coleta e classificação dos textos opinativos gerados por estes usuários, uma vez que tais informações refletem a qualidade do produto ou serviço prestado. O presente trabalho demonstra a aplicação de uma técnica de mineração de opinião que consiste em classificar críticas sobre cinema em positivas, negativas ou neutras, de acordo com o sentimento empregado. Além disso, são apresentados os conceitos gerais da metodologia que rege este trabalho, bem como as limitações da pesquisa e os possíveis campos para trabalho futuro.

Palavras-chave: Mineração de opinião. Classificação de críticas. Análise de sentimento.

ABSTRACT

In the recent world stage, marked by globalization and the constant competitivity in the companies, the great amount of data generated daily on the web includes an important source opinion about varied topics, including review products and services availables in the market. In this context, arise the need for automating the collection process and text classification opinionated generated by web users, because these informations reveal the quality of the product or service provided. This paper presents the application of an opinion mining technique to classify comments in portuguese about cinema in positive, negative or neutral, based on sentiment expressed in the text. Moreover, are presented the general methodology concepts of this study, as well as the search limitations and the possible fields for future study.

Keywords: Opinion mining. Review classification. Sentiment analysis.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Com a popularização da internet no início do século XXI, as relações comerciais e as comunicações foram revolucionadas e adquiriram dimensões globais. Com esta facilidade e rapidez de interação com o mundo, muitas empresas começaram a se informatizar e, neste processo de conexão com a web, acabaram por divulgar seus serviços, produtos e marcas para todo o planeta. Desta forma, enquanto potencializam seu poder de negócio, tornam a pesquisa de mercado dos usuários mais simplificada.

Neste cenário digital, as pessoas passaram a se interconectar e, atualmente, compartilham informações na web por meio de blogs, redes sociais, fóruns e outros servidores de comunicação, disseminando dados e repassando notícias para o mundo inteiro de forma praticamente imediata.

No mercado competitivo dos últimos anos, as empresas passaram a buscar estratégias para se destacarem, sendo que a principal delas está associada à satisfação do usuário final. A internet, por ser a mais rica fonte de comentários e opiniões de consumidores, tem sido o principal alvo de análise das grandes empresas.

Com isso, a opinião dos internautas se tornou objeto de muito interesse por parte das empresas, afinal o comentário do consumidor representa um termômetro que indica o quanto o produto e/ou serviço está agradando e em quais pontos precisa melhorar.

Por se tratar de uma quantidade numerosa de informação, o processo de filtragem e classificação de opiniões da web representa um trabalho árduo, que demanda muito tempo e que exige múltiplas pessoas envolvidas, dependendo da popularidade do objeto pesquisado.

Para automatizar este processo, muito se tem explorado acerca do processamento de linguagem natural, um dos vastos campos de estudo da Inteligência Artificial. O histórico de seus experimentos aponta que sua aplicação inicial revela certo grau de complexidade, mas retorna resultados satisfatórios quando se utilizam métodos apropriados. (SILVA et al., 2007; CRUZ et al., 2008).

O processamento de linguagem natural trabalha com a linguística humana e se concentra em analisar textos escritos em busca de um resultado específico

referente ao seu conteúdo, de forma que procura palavras e combinações linguísticas que indiquem algo relevante para sua classificação ou descoberta do sentido principal. (SILVA et al., 2007).

Uma das aplicações desta área é a classificação de documentos, que busca descobrir qual o tipo do documento tratado (CRUZ et al., 2008). A principal utilidade desta ferramenta é a filtragem de textos, pois em muitos casos é necessário saber se o texto segue o foco estudado, para que somente assim seja considerado nas demais análises. Por exemplo, se o objetivo for encontrar o assunto principal de reportagens jornalísticas, primeiramente será necessário confirmar se um dado documento realmente se trata de uma reportagem.

No ramo da publicidade, utilizam-se sistemas de linguagem natural capazes de detectar o assunto de conversas on-line, pois ao desvendar os interesses de um usuário, será possível oferecer-lhe os produtos mais adequados através de propagandas virtuais.

O presente trabalho teve seu foco de estudo fundamentado em outra aplicação do processamento linguístico, que possui certa semelhança com a classificação de documentos, mas que se concentra em detectar opiniões/sentimentos expressos no comentário de um autor. Trata-se da mineração de opinião, que procura, principalmente, descobrir o nível de satisfação de um autor em relação a determinado alvo de crítica, que pode ser uma empresa, produto, serviço, pessoa ou qualquer outro tipo de objeto passível de julgamento. Para o estudo em questão, serão classificados comentários em português sobre cinema.

A ideia básica deste processo é definir se o comentário mostra um sentimento positivo, neutro ou negativo em relação ao objeto criticado, e procura determinar qual o grau deste sentimento, ou seja, o peso positivo ou negativo que está inserido no comentário.

O interesse pelo campo de estudo da mineração de opinião surgiu, primeiramente, pela potencialidade que esta área demonstra para contribuir com a evolução do processamento de informações e também devido à escassa quantidade de trabalhos e artigos publicados com foco neste estudo, sendo que desta quantidade mínima uma baixa porcentagem está na língua portuguesa.

Esta área se mostra promissora ao possibilitar a execução de uma tarefa complexa por meio de uma solução computacional e, portanto, automática, que não

se resume somente a uma lógica objetiva, mas que exige certa subjetividade ao lidar com a interpretação da linguagem humana.

A baixa quantidade de estudos publicados sobre este assunto até o momento certamente se deve por tratar de uma área com pouco tempo de surgimento, que ainda não foi tão popularizada no meio acadêmico e que, por conta disso, está em processo de evolução, embora existam pesquisadores mais renomados como Turney (2002) que já tenham elaborado experimentos mais complexos sobre o assunto. Além disso, a maioria destes estudos está catalogada na língua inglesa e, do restante, uma vasta porcentagem se encontra em espanhol - como é o caso do artigo publicado por Cruz et al. (2008).

Para a língua portuguesa, a quantidade de publicações é mínima e, assim, o presente trabalho se apresenta em contribuição ao meio acadêmico do idioma, seja para revelar a metodologia adotada como satisfatória ou até mesmo para mostrar sua ineficiência, de acordo com os resultados obtidos. Desta forma, os experimentos irão demonstrar se a técnica empregada é eficiente a ponto de ser tomada como base para trabalhos futuros.

A escolha de cinema como gênero para o corpus de mineração de opinião se originou por tratar de um assunto de entretenimento e, com isso, de grande interesse do público em geral. Outro fator atrativo deste campo é a rica diversidade existente entre os filmes, com seus estilos variados (comédia, terror, drama,...) e diversos quesitos integrantes, como o roteiro, os efeitos especiais, a incorporação dos atores em seus personagens, o jogo de câmeras, entre outros. Isso garante maior variação de adjetivos nos comentários e focos diferentes para cada crítica, o que torna a sua classificação ainda mais desafiadora e interessante.

Um sistema de mineração de opiniões eficiente pode representar uma ferramenta de grande valia, talvez capaz de decidir o destino de uma empresa entre o sucesso e o fracasso, uma vez que as empresas teriam conhecimento das potencialidades a serem mantidas em seus produtos bem como dos pontos que poderão levá-la a um possível declínio. Esta situação também se encaixa na campanha política de um candidato à eleição, na popularidade de uma banda/grupo musical ou até mesmo na reputação de atores e artistas que se preocupem com a imagem pessoal. Para os produtores de filmes, ter conhecimento dos detalhes que agradaram o público potencializa a produção dos próximos filmes e, até mesmo, pode indicar se vale a pena produzir uma continuação do filme atual.

Com a posse de informações importantes obtidas de forma tão simplificada, as estratégias seriam pensadas com maior facilidade, gerando resultados de melhor qualidade que provavelmente decidirão o seu sucesso.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Processar comentários em português sobre cinema com a intenção de coletar o sentimento empregado pelo autor, medindo sua carga de satisfação e insatisfação para, assim, classificá-lo como neutro, positivo ou negativo, através da aplicação de uma técnica linguística em um corpus montado a partir da coleta de comentários da web.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar pesquisa bibliográfica a fim de coletar informações relevantes e expandir o conhecimento sobre o tema proposto para desenvolver um trabalho de qualidade;
- Montar um corpus de comentários em português sobre cinema, usando como fonte de informação o site adorocinema.com;
- Analisar a combinação dos tipos de palavras em comentários sobre cinema, a fim de encontrar padrões que facilitem a classificação do sentimento empregado;
- Estabelecer regras baseadas na gramática para a criação de uma técnica satisfatória de análise de opinião;
- Capacitar um sistema a medir a carga positiva e negativa dos comentários, baseando-se em dois conjuntos de palavras raiz, um positivo e um negativo;
- Classificar determinado comentário como neutro, positivo ou negativo;
- Verificar se a técnica aplicada possibilita a construção de um sistema de processamento de opinião eficiente.

3 INTELIGÊNCIA ARTIFICIAL

Um dos campos de estudo mais amplos e complexos de Ciência da Computação é a Inteligência Artificial (IA). Esta área se debruça na criação de sistemas considerados inteligentes, uma vez que buscam imitar o comportamento inteligente (baseando-se, muitas vezes, nos seres vivos) e encontrar soluções para problemas que exijam raciocínio. (SATO, 2009). Para Charniak e Mcdermott, que foram citados por Russel e Norvig (2004, p. 15), IA representa “o estudo das faculdades mentais pelo uso de modelos computacionais”. De forma prática, Kurzweil, que também foi citado por Russel e Norvig (2004, p. 15), considera IA como “a arte de criar máquinas que executam funções que exigem inteligência [...]”.

Os sistemas inteligentes utilizam as mesmas linguagens computacionais de programas convencionais, mas as grandes diferenças estão na lógica e no processamento de informações. Alguns destes sistemas oferecem um funcionamento simples, como a indicação de uma solução x para um determinado problema y, enquanto que outros tentam simular o funcionamento dos neurônios humanos, combinando informações e utilizando conhecimento prévio para apresentar a solução mais adequada possível. (SATO, 2009).

3.1 APLICAÇÕES DE IA

Pela sua grandeza, a Inteligência Artificial se ramifica em vários estudos específicos do comportamento humano, sendo que cada um se concentra em reproduzir uma habilidade diferente. Como exemplo, podem ser citados os sistemas que trabalham com o reconhecimento de fala, pois interpretam palavras proferidas pelo usuário através das ondas sonoras captadas. (PETRY, 2008). Neste processo, podem ocorrer muitas variações na fala e, por este motivo, os sistemas deste tipo necessitam de conhecimento prévio sobre a estrutura do idioma empregado, como a formação de palavras e as combinações lógicas possíveis entre elas para formar frases.

Alguns sistemas de voz, além de cumprir sua tarefa fundamental, também detectam o idioma utilizado, como é o caso do reconhecimento de fala do Google, e assim se tornam ainda mais complexos e exigem um banco de conhecimento mais extenso.

A Inteligência Artificial também marca forte presença na indústria robótica, pois muitos robôs são projetados para agir e se movimentar como o homem, inclusive imitando sua estrutura física. Existem robôs capazes de movimentar objetos, fazer faxina, conversar e, até mesmo, praticar esportes, entre outras atividades.

Nos jogos eletrônicos, os personagens que são controlados pela máquina requerem uma programação inteligente, de maneira que possam executar ações coerentes e, de preferência, imprevisíveis para cada situação. Em muitos casos, precisam ser capazes de bolar estratégias para tornar o jogo ainda mais desafiador, como é o caso dos simuladores de esporte.

3.1.1 Aplicações de IA no processamento de linguagem natural

Para a área linguística, também existe grande contribuição dos sistemas inteligentes, uma vez que a análise textual se apresenta, em muitas situações, como uma tarefa complexa até mesmo para um ser humano, ao envolver um grande número de regras gramaticais e exigir vasto conhecimento linguístico. Sistemas que processam este tipo de informação são muito abrangentes, pois podem atender a qualquer tipo de usuário computacional, desde o mais iniciante ao mais avançado.

Um exemplo deste tipo de sistema são os corretores ortográficos, que em maior parte estão presentes nos editores de texto, como o famoso Microsoft Word. Conforme o usuário digita seu texto, o sistema analisa cada palavra para detectar um possível erro de gramática, enquanto que também processa a combinação das palavras dispostas em frase para aplicar uma série de verificações, como por exemplo, se estão corretamente empregadas, se transmitem coesão e se possuem concordância. (MEDEIROS, 1995). Além disso, em alguns sistemas (como é o caso do Word) são sugeridas possíveis correções para o suposto erro encontrado.

Os tradutores de idioma são mais um exemplo de sistema linguístico de grande utilidade, seja para a leitura de um documento em um idioma desconhecido pelo usuário ou até mesmo para o aprendizado. Os mais sofisticados suportam uma gama variada de idiomas, como é o caso do Google Tradutor.

Algoritmos processadores de texto são empregados até mesmo nos softwares desenvolvedores de sistemas, que é o caso dos compiladores, uma ferramenta que processa linhas de comando de determinada linguagem de programação e realiza

análises léxica, sintática e semântica e, seguindo a lógica empregada, gera um código objeto, sendo geralmente um sistema ou programa executável. (GUERBER, 2007).

Existem sistemas de IA voltados para a gestão comercial, que auxiliam empresas a divulgar seus produtos para as pessoas certas ou que até mesmo coletam dados importantes para a sua permanência no mercado. Partindo da ideia que a propaganda pode ser uma ferramenta muito eficaz para impulsionar o consumo, constantes estudos são realizados na intenção de criar sistemas capazes de coletar dados transmitidos por usuários na internet (como conversas em redes sociais e pesquisas na web) para tentar descobrir seus interesses pessoais e, com isso, poder oferecer-lhes os produtos mais apropriados. Desta forma, a tendência é que a empresa obtenha um resultado melhorado de vendas.

3.1.2 Aplicação de IA na mineração de opinião

Outro fator importante a ser observado pelas empresas está ligado à satisfação dos consumidores, uma vez que usuários mais satisfeitos tendem a continuar com os mesmos produtos/serviços. Com a chegada da internet, a comunicação explorou alcances globais, e o compartilhamento de informações ficou muito mais simplificado: qualquer pessoa do planeta que esteja conectada à Internet pode acessar um comentário público expresso por algum usuário na web.

Neste contexto, tendo em vista a facilidade de acesso e a propagação da informação, a internet se consolidou como uma das mais ricas fontes de informação da atualidade e, por isso, se tornou alvo de grande atenção por parte das empresas. Por meio de blogs, redes sociais, fóruns e demais servidores de comunicação os usuários publicam comentários na web sobre os mais variados assuntos, incluindo opiniões pessoais sobre empresas e produtos. (GUEDES; AFONSO; MAGALHÃES, 2010).

Com a posse de críticas valiosas sobre seus produtos, uma empresa adquire um conhecimento maior dos seus pontos fortes, para poder explorá-los, bem como passam a observar suas fragilidades, que servem como um alerta para algo que precisa ser melhorado.

A grande dificuldade se concentra, primeiramente, na coleta dessas informações, visto que o território da web é imenso e a variedade de dados é

incalculável. Seria necessária uma análise exaustiva para a filtragem dos dados, que demandaria tempo e trabalho diário excessivo para ser realizado por pessoas. Além disso, as críticas precisariam ser analisadas individualmente (com o objetivo de extrair a opinião de cada uma) e, assim, considerando que um resultado eficiente dependeria de uma quantidade razoável de comentários, este processo seria muito trabalhoso e demorado.

Para superar os desafios desta tarefa, surgem os estudos sobre a mineração de opinião, que propõem métodos para a automatização inteligente da análise crítica do comentário. (SANTOS, 2010). É importante destacar que a extração de comentários da web relacionados ao tema específico se trata de um processo separado, e seria realizado por outro sistema, senão manualmente.

A mineração de opinião consiste basicamente na classificação de comentários sobre um assunto específico (como esportes, cinema, celebridades ou produtos/serviços de uma empresa) a fim de determinar se a opinião do autor assume uma posição positiva, negativa ou neutra sobre o objeto em discussão. (SANTOS, 2010).

No presente trabalho, serão classificadas críticas em português sobre cinema através da aplicação de métodos específicos que se baseiam nos estudos sobre mineração de opinião, uma ramificação da Inteligência Artificial que utiliza o conceito de Processamento de Linguagem Natural (PLN), uma vez que trabalha com a análise da linguagem humana para o reconhecimento de informações. (SILVA et al., 2007).

4 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O surgimento dos computadores, além de permitir avanços consideráveis nos diversos campos do estudo científico, também abriu caminho para novas frentes de pesquisa que, até então, não eram sequer cogitadas. Entre elas, podem-se destacar os estudos sobre PLN, que vêm se desenvolvendo conforme ocorre o aperfeiçoamento da comunicação entre homem e máquina. (SILVA et al., 2007).

O primeiro grande desafio era fazer a máquina compreender instruções para a execução de tarefas. Foi então que surgiram as primeiras linguagens de programação, que permitiram a comunicação homem-máquina através da confusa linguagem de máquina. (SILVA et al., 2007).

Cada vez mais, as linguagens de programação foram aprimoradas com a intenção de se aproximarem da linguagem humana e, assim, tornar esta comunicação mais simples e atraente para os usuários. Embora muitas linguagens de programação tenham adquirido níveis bastante inteligíveis, ainda se distanciam da linguagem natural, já que, entre outros motivos, não são capazes de interpretar instruções que não estejam na sintaxe exata proposta pela linguagem. (SILVA et al., 2007).

Para otimizar ainda mais esta interação do homem com o computador, muitos pesquisadores estudaram a criação de interfaces gráficas, que permitiram ocultar a maneira primitiva de comunicação, mascarando-a com objetos gráficos mais agradáveis e aproximados da leitura humana. (SILVA et al., 2007).

Apesar de tudo, os projetos mais ambiciosos desta área continuam sendo a criação de sistemas capazes de interpretar e gerar mensagens codificadas em línguas naturais, que tornem possível uma interação verbal do homem com a máquina. (SILVA et al., 2007).

4.1 SISTEMAS DE PLN

Os sistemas que trabalham com PLN podem dispor de duas funções principais: a interpretação e a geração de língua natural. Uma pode ser tão complexa quanto a outra, dependendo do método a ser utilizado e da aplicação específica do sistema. (SILVA et al., 2007). Os corretores ortográficos, por exemplo,

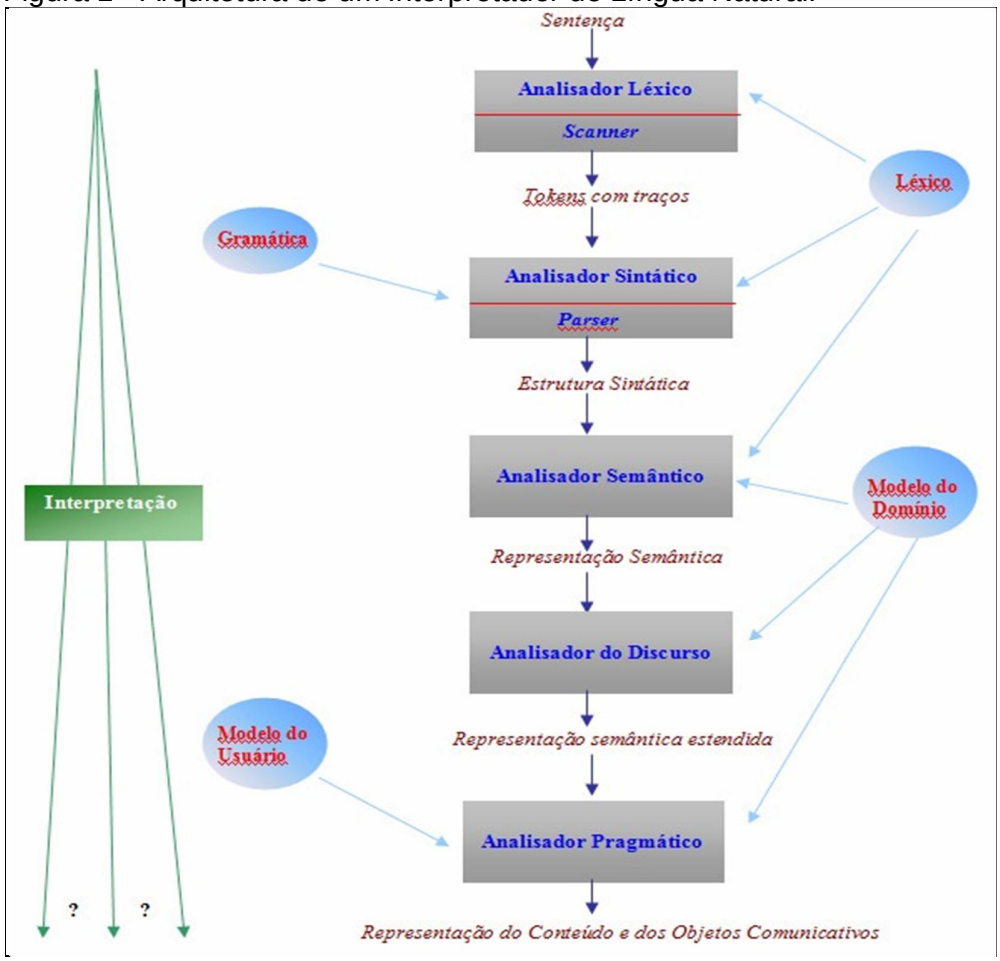
utilizam apenas a primeira função, enquanto que os sumarizadores e os tradutores de texto necessitam da combinação das duas tarefas.

4.1.1 Interpretador de língua natural

Um interpretador realiza diferentes níveis de processamento linguístico de forma a coletar descrições específicas das palavras em cada etapa. Cada uma destas informações é armazenada no léxico, que "[...] consiste em um conjunto de palavras ou expressões da língua associadas a um conjunto de atributos [...]". (SILVA et al., 2007, p. 33). Portanto, o léxico se trata de um fundamental recurso para o PLN, uma vez que as fases de interpretação e geração do código em língua natural dependem do acesso e manipulação de suas informações.

A arquitetura genérica de um sistema interpretador de LN é mostrada na Figura 1.

Figura 1 - Arquitetura de um Interpretador de Língua Natural.



Fonte: Silva et al. (2007).

Em geral, um interpretador pode iniciar o processamento do texto com uma análise morfológica. A Morfologia estuda a estrutura, formação e classificação das palavras, tratando-as individualmente, sem considerar sua participação na frase ou texto. Portanto, esta etapa se encarrega de identificar os morfemas (unidade mínima dotada de significado na linguagem), que podem ser do tipo gramatical – tais como os marcadores de flexão das palavras, que identificam traços nominais como gênero e número, e traços verbais como pessoa, número e tempo – ou do tipo lexical, quando se ocupam do processo de derivação das palavras. (ROSA, 2005).

Exemplificando casos de morfema do tipo gramatical na língua portuguesa, a palavra “cadeiras” possui uma flexão de número (morfema “s”) em relação à palavra “cadeira”, pois está no plural; a palavra “poderemos” possui uma flexão (morfema “emos”) de pessoa (1ª pessoa), número (plural) e tempo (futuro). Agora, como exemplo de morfema do tipo lexical, dada uma palavra base (radical), são acrescentados sufixos (sucedem o radical) e/ou prefixos (antecedem o radical) para formar uma nova palavra da língua: a palavra “pedreiro” é uma derivação da palavra “pedra”, onde foi acrescentado o sufixo “eiro”; a palavra “insuficiente” é uma derivação da palavra “suficiente” pela adição do prefixo “in”, que indica negação. (ROSA, 2005).

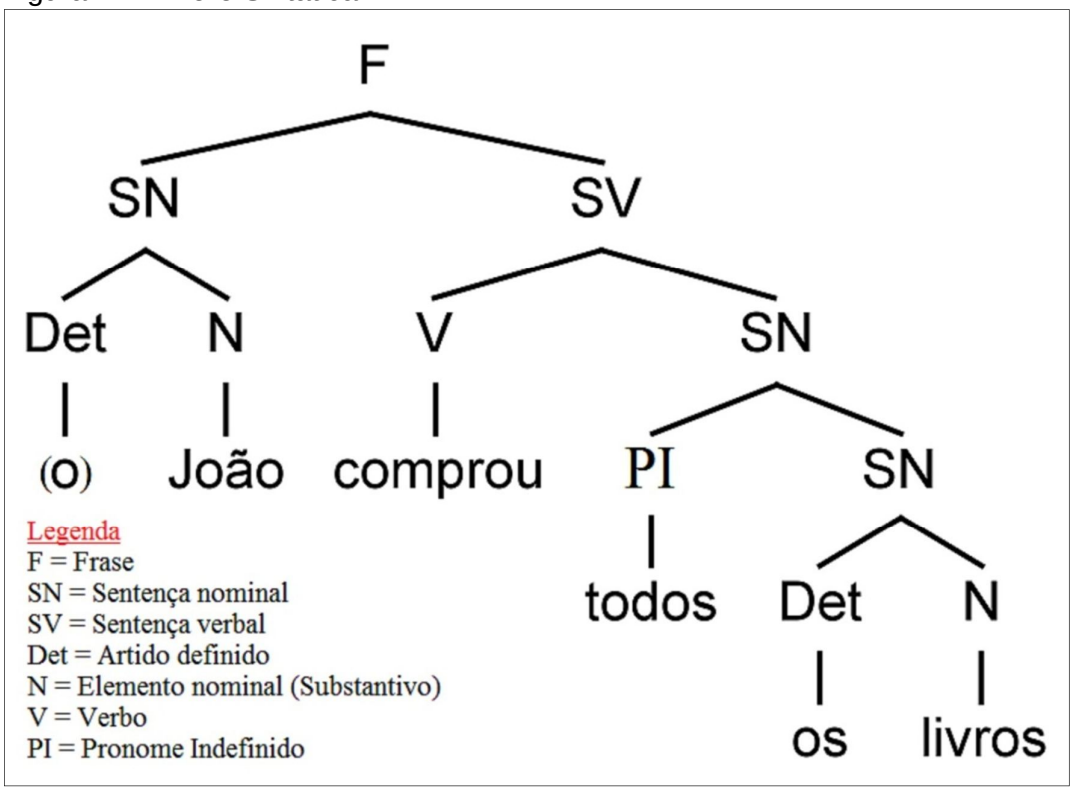
Durante a análise morfológica, são determinadas as relações que cada morfema assumirá na definição da palavra e, assim, poderá definir fenômenos entre palavras, como concordâncias verbal e nominal. Esta etapa é facultativa, podendo ser realizada dependendo da estrutura do léxico e dos atributos exigidos pela aplicação. (SILVA et al., 2007).

Como exemplo de concordância verbal, a locução “nós poderemos” possui concordância, pois o pronome “nós” exige um verbo flexionado na 3ª pessoa do plural (como aconteceu com a palavra “poderemos”, neste caso). Para exemplificar concordância nominal, a expressão “os policiais” está em concordância, pois o substantivo “policiais” está no plural e, assim, exige um artigo também no plural. (ROSA, 2005).

Em paralelo com a etapa morfológica, ou até mesmo após ela, deve ser executada uma análise léxica (ou scan), que identifica os componentes da sentença e os separa em cadeias unitárias, conhecidas como tokens, associando a cada um deles os atributos gramaticais e semânticos encontrados no conjunto léxico. (SILVA et al., 2007).

Após o tratamento de cada palavra e símbolo como um token, ocorre o processo de análise sintática (ou parse), que se encarrega de construir uma estrutura sintática válida para a sentença em análise. Geralmente, utiliza-se uma representação gramatical parcial da língua natural em questão que envolve apenas as construções de interesse para a aplicação, com o intuito de manter a eficiência do processo. (SILVA et al., 2007). A Figura 2 ilustra esta etapa, representando uma árvore sintática gerada para a frase “João comprou todos os livros”.

Figura 2 – Árvore Sintática.



Fonte: Hefren (2010).

Enquanto a análise sintática se especializa em organizar os elementos de ordem linguística dos componentes da sentença, o analisador semântico, que representa a próxima etapa, é responsável por identificar a relação entre os componentes da sentença em nível de significado, fornecendo a interpretação geral do texto. Esta etapa se baseia na estrutura sintática válida gerada pela fase anterior para relacionar componentes, por exemplo, durante a detecção do sentido real de palavras com mais de um significado, através do domínio reconhecido no contexto. (SILVA et al., 2007).

Caso o texto de entrada seja um discurso multisentencial, surge uma dificuldade ainda maior se comparado à análise de uma sentença individual. Isso porque, neste tipo de discurso, é muito comum que o significado de uma sentença dependa de sentenças anteriores e, da mesma forma, que influencie no significado de sentenças seguintes. (SILVA et al., 2007). Por exemplo, frequentemente utilizam-se referências anafóricas, que servem para retomar algo ou alguém já mencionado anteriormente no texto, como utilizar a expressão “o garoto” referindo-se a “Joaquim”. O analisador de discurso trata exatamente este tipo de relação, estendendo a representação semântica com as funções das figuras de discurso.

A interpretação da mensagem original de uma sentença pode, ainda, depender de aspectos pragmáticos da comunicação, ou seja, a significância prática e objetiva das expressões, como o exemplo da frase interrogativa “Você sabe que horas são?”, que pode ser interpretada como a solicitação de uma resposta ou até mesmo como a imposição de autoridade sobre um atraso ocorrido. Para tratar casos deste tipo, deve ser levado em conta o contexto de ocorrência do discurso, a fim de evitar interpretações errôneas de sentido. Este processo representa a última etapa de um interpretador, conhecida como análise pragmática. (SILVA et al., 2007).

É importante ressaltar que os cinco processos (léxico, sintático, semântico, discursivo e pragmático) não são, necessariamente, executados nesta sequência, podendo haver processos combinados de forma distinta dependendo da especificação do projeto que está sendo desenvolvido. (SILVA et al., 2007).

4.1.2 Gerador de língua natural

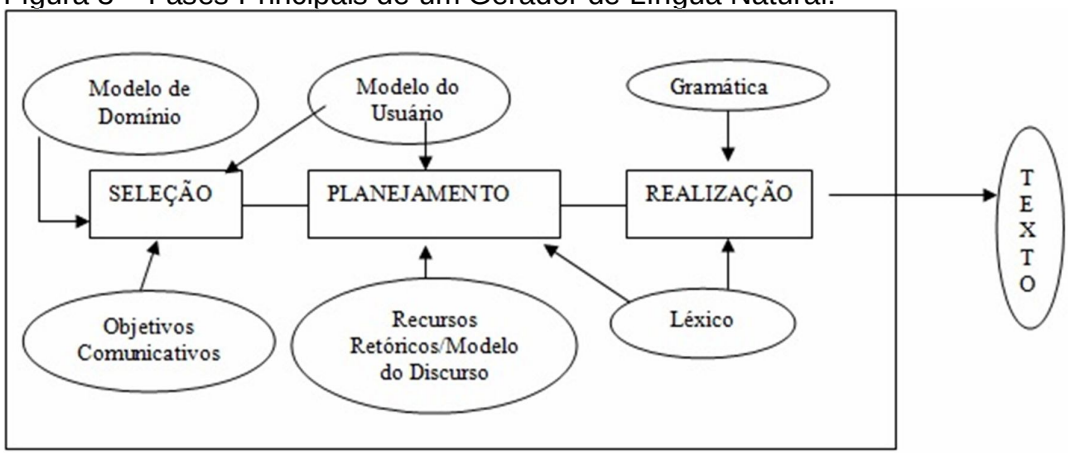
A função do gerador de língua natural consiste basicamente em construir uma forma textual em linguagem humana, tomando parte de elementos conceituais e de um objetivo específico de comunicação. A arquitetura mais comum de um gerador inclui três etapas: a fase de seleção do conteúdo, o planejamento da estrutura textual e a aplicação desta estrutura na criação do texto. (SILVA et al., 2007).

Da mesma forma que no interpretador, são utilizados recursos léxicos e uma gramática determinada, desta vez para selecionar os componentes que farão parte do texto. Para a estruturação e construção do texto também são exigidos certos conhecimentos linguísticos, que permitam encontrar a combinação ideal entre os

conceitos discursivos a fim de atingir o objetivo comunicativo desejado. (SILVA et al., 2007).

As fases principais de um sistema gerador de LN são apresentadas na Figura 3.

Figura 3 – Fases Principais de um Gerador de Língua Natural.



Fonte: Silva et al. (2007).

4.1.3 Tagger

O processo de identificar a classe gramatical das palavras de uma sentença pode ser executado por um tipo de aplicativo conhecido como Tagger. O nome deste tipo de aplicativo origina-se do conceito de tag, que representa um termo associado a determinado conteúdo (como uma imagem, documento ou música), que serve para identificar o seu gênero contextual facilitando, assim, a busca de conteúdos por tema. Em outras palavras, sua função se relaciona com a tradução da palavra vinda do inglês – “etiqueta”. Portanto, estas palavras-chave permitem uma maior organização das informações na Internet de maneira que informações relacionadas sejam agrupadas. (ASSIS, 2009).

Partindo da tradução da palavra tag e seguindo padrões da língua inglesa, tagger pode ser definido como um “etiquetador”, responsável por determinar atributos específicos para cada componente de um universo. No contexto presente, ele se encarrega de detectar a classe gramatical de cada palavra em uma sentença, como adjetivos, pronomes, substantivos, entre outras classificações.

Segundo o desenvolvedor de um sistema tagger conhecido como Tree-Tagger, este tipo de sistema representa um interpretador morfossintático.

(GAMALLO, 2005). Desta maneira, pode-se concluir que este interpretador realiza apenas as três primeiras etapas de interpretação de LN (as análises: morfológica, léxica e sintática), visto que não necessita considerar a significação semântica das palavras.

Com o conhecimento da classe gramatical de cada palavra que compõe a sentença que, no caso, será um comentário sobre cinema, será possível coletar os bigramas relevantes para a detecção da opinião expressa no comentário. As regras que definem a relevância dos bigramas serão apresentadas na seção Metodologia deste trabalho.

Cada bigrama que se encaixar nas condições impostas será processado com a finalidade de se calcular a sua orientação semântica.

4.2 ORIENTAÇÃO SEMÂNTICA

A orientação semântica de uma palavra ou expressão consiste em definir se ela possui uma conotação positiva ou negativa na frase em termos de sentido qualitativo. (CRUZ et al., 2008). Por exemplo, o termo “ótimo” referencia algo bom, atribuindo um sentido positivo ao objeto abordado na frase. Enquanto isso, a expressão “mal” indica um valor negativo para algo que certamente não agradou o autor do comentário.

No seu conceito absoluto, a orientação semântica se trata de um valor matemático, da classe dos reais, que representa a medida subjetiva do sentido de uma palavra ou expressão. Assim, possuindo um valor positivo (maior que zero), possui implicações positivas e, da mesma forma, se apresentar valor negativo (menor que zero), indica um sentimento negativo. (CRUZ et al., 2008). Vale destacar que esta medida, além de indicar a conotação mais aproximada da palavra, também indica o grau de intensidade do seu sentido. Logo, altos valores positivos estão relacionados com designações ótimas, enquanto valores extremamente negativos representam orientações de sentido péssimo para a palavra. Valores muito próximos do marco zero atribuem um sentido duvidoso para a palavra e, nestas situações, não se deve levar em conta o sinal positivo ou negativo da medida semântica. Assim, o sentido de palavras com tais valores talvez seja mais bem retratado como neutro. O buscador Bing oferece um operador, denominado NEAR, que pode ser utilizado por

sistemas de PLN no cálculo da orientação semântica. Tal operador será apresentado no tópico seguinte.

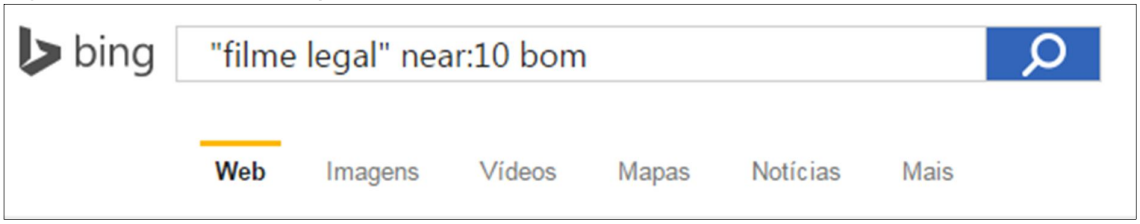
4.2.1 Operador NEAR

O buscador Bing, assim como vários outros sistemas de busca da web, possui operadores especiais que podem ser utilizados para restringir o efeito da busca.

O comando NEAR (traduzido para o português como “próximo”), que será utilizado no presente trabalho, se trata de um operador de proximidade que impõe um limite na busca de dois ou mais termos (palavras ou expressões) em um mesmo documento, baseando-se em uma quantidade máxima de palavras que podem intermediá-los. (DICAS..., c1996-2014). Desta forma, permite procurar por conteúdos que contenham os dois termos pesquisados sem estarem necessariamente ligados por alguma hierarquia ou dependência contextual, ao contrário do que ocorreria em uma busca comum.

A sintaxe do operador NEAR é apresentada na Figura 4. Após o símbolo de dois pontos, deve ser indicada a quantidade máxima de palavras que pode separar os dois termos. (DICAS..., c1996-2014).

Figura 4 – Sintaxe do Operador NEAR.



Fonte: Elaborada pelo autor.

Como pode ser visualizado na Figura 4, o primeiro termo que, neste caso, será o bigrama analisado, é colocado entre aspas a fim de se preservar a sequência de suas palavras, de maneira que a busca retorne apenas resultados da palavra “filme” seguida por “legal”, mas não o contrário.

A ordem das palavras-chave no comando não é um fator relevante para a consulta e, desta maneira, pesquisar no Bing “ciência near:10 computação” poderá trazer resultados com ordens diferentes de ocorrência dos termos, como “[...] ciência

da computação [...]” e “A computação é a ciência que [...]”. Ou seja, a ordem dos termos nos conteúdos encontrados não seguirá a ordem digitada no comando NEAR.

A utilização deste comando, que influenciará diretamente no resultado de classificação opinativa do comentário, representa um dos processos da técnica que rege este atual projeto.

5 MINERAÇÃO DE DADOS

Em paralelo com o crescimento tecnológico e a popularização da internet nos últimos anos, expandiu-se de maneira semelhante a produção de informação virtual, em geral contida na web. Este acúmulo de dados acontece diariamente e em proporções grandiosas e, à medida que a computação avança, a tendência é que este crescimento ocorra em níveis ainda maiores.

Em meio a este emaranhado de informações dos mais variados tipos e assuntos, surge a necessidade de filtrar os dados na intenção de se extrair conteúdos relevantes para o aperfeiçoamento de algum tipo de atividade.

A partir deste ponto, surge o conceito de mineração de dados, ou data mining, que representa a técnica de extração ou filtragem de dados úteis em meio a vastos volumes de informação. (HAN; KAMBER, 2006). Em outras palavras, designa a área de estudo da Inteligência Artificial responsável por obter conhecimento, desvendar estruturas e extrair padrões ocultos em massas de dados. (GUEDES; AFONSO; MAGALHÃES, 2010).

Ao analisar conjuntos de dados diversos, como aqueles dispostos na web, por exemplo, é possível buscar padrões implícitos de informação que seja relevante para as mais variadas aplicações comerciais, como para traçar o perfil dos consumidores de determinado produto, ampliar um plano de negócios, analisar riscos e otimizar técnicas de divulgação e propaganda.

As etapas da técnica de mineração de dados abrangem a seleção de conteúdos apropriados para a coleta de informação útil e, para esta decisão, deve-se levar em conta fatores como a confiabilidade e o contexto principal do conjunto de dados. Em seguida, é realizada uma breve checagem dos dados a fim de coletar possíveis inconsistências, que deverão ser prontamente corrigidas para assegurar a qualidade dos conhecimentos a serem extraídos nas etapas posteriores.

Os dados são analisados e adaptados para um modelo que pode ser interpretado por um algoritmo de mineração que, baseado em critérios pré-definidos, converte os dados gerais em informações úteis, que podem revelar padrões e relacionar novos conhecimentos.

Os algoritmos empregados nas técnicas de mineração de dados se baseiam na aprendizagem ou na classificação através da implementação de redes neurais e métodos estatísticos. Os resultados geralmente retornam regras e hipóteses.

5.1 MINERAÇÃO DE OPINIÃO

A mineração de opinião, ou opinion mining, também conhecida como análise de sentimento, representa um dos ramos de estudo do PLN e uma derivação da mineração de dados (data mining), e se concentra basicamente em extrair o sentimento opinativo empregado em um texto que seja direcionado à crítica de algum objeto passível de qualificação. (SANTOS, 2010). O alvo pode ser um produto ou aparelho, a campanha política de um candidato à eleição ou até mesmo conteúdos culturais, como músicas, livros e filmes.

Os estudos desta área se baseiam em classificar como positivo ou negativo o sentimento expresso em um texto crítico, com a intenção de indicar se o autor assume uma posição a favor ou contra o objeto em discussão, de acordo com a polaridade de significado das palavras contidas no texto.

Muitas vezes considerada como uma derivação da tarefa de mineração de documentos, que consiste em identificar o caráter temático de um documento (CRUZ et al., 2008), a análise de sentimento se diferencia pela natureza subjetiva do conteúdo que precisa ser interpretado, visto que, em muitas ocasiões, necessita de conhecimento prévio sobre o assunto, além de considerar fatores linguísticos como a sintaxe e a semântica das palavras.

Por exemplo, em uma crítica onde o autor declara: “Este filme apresenta mais uma das histórias empolgantes inventadas pelo diretor”, existem duas interpretações possíveis. A primeira, levada ao pé da letra, indica que o autor aprovou a história contada no filme, enquanto uma segunda interpretação, levando em consideração um tom irônico, revela certo grau de frustração por mais um filme monótono produzido pelo diretor em questão.

Neste sentido, uma interpretação segura do comentário requer conhecimentos sobre quem é o diretor do filme, quais os demais filmes produzidos por ele e se estes filmes são considerados bons.

Por trabalhar, em grande parte, com conteúdos textuais gerados por usuários da internet, a linguagem rebuscada nos comentários torna-se mais um fator de dificuldade para a análise de opinião. Para resolver esta questão, um prévio tratamento do texto através de um corretor ortográfico seria bastante recomendado. Além disso, o processo de mineração contemplaria maior efeito se considerasse gírias durante a análise da opinião.

Classificar de forma coerente o sentimento expresso em um comentário depende diretamente da aplicação de uma técnica eficiente de análise, mas também pode sofrer variações de acordo com as características do texto. Em casos com muita variação de humor, por exemplo, seria mais difícil decidir a tendência opinativa do autor. Assim, surge uma nova classificação de crítica, que pode ser definida como uma opinião neutra.

Uma opinião também pode ser interpretada como neutra quando o autor não expressar nenhum termo que demonstre uma posição opinativa. Geralmente, isso ocorre quando o texto trata de descrever características do objeto. No comentário “Este filme foi filmado no Brasil!”, não existe nenhuma referência crítica, mas apenas foi informado o lugar onde o filme foi gravado.

Por conta de sua grande proximidade para com o mundo real, a mineração de opinião vem se consolidando como uma área de grande interesse acadêmico, tanto por parte de novos ingressantes na área de linguagem natural, bem como de mestres e doutores, que buscam aprimorar os seus conhecimentos e contribuir para o crescimento desta área tão promissora.

6 METODOLOGIA

O presente trabalho de mineração de opinião se concentrou em identificar a opinião crítica de comentários em português sobre cinema, classificando-os como positivo, negativo ou neutro de acordo com a orientação semântica expressada no texto.

Para tanto, inicialmente foi montado um corpus de comentários em português sobre cinema a partir da coleta manual de críticas existentes no site Adorocinema.com. Foram coletados os comentários que apresentaram boa linguagem e coerência opinativa. De cada comentário, foram extraídos bigramas relevantes, que eram assim considerados ao atender a uma das sete regras estabelecidas – estas regras estão relacionadas com a combinação de classes gramaticais das duas palavras que formam o bigrama. Para cada bigrama relevante foi calculada a sua orientação semântica através de um método estatístico definido pelo algoritmo PMI-IR, que será detalhado mais adiante.

Os valores semânticos obtidos para cada bigrama de uma crítica foram somados entre si para determinar a classificação final do comentário: para valores menores que zero, o comentário foi classificado como negativo. Do contrário, foi contado como crítica positiva. Considerando uma pequena margem de erro, valores muito próximos da marca zero, fossem positivos ou negativos, implicaram numa classificação neutra para o comentário. Os tópicos seguintes abordam cada um dos passos citados.

6.1 OBTENÇÃO E MONTAGEM DO CORPUS DE COMENTÁRIOS

A técnica de mineração de opinião exposta neste trabalho foi aplicada em um corpus de comentários sobre cinema. Tais comentários foram extraídos do site AdoroCinema.com, que foi escolhido pela qualidade do seu conteúdo jornalístico, fato que revela a confiabilidade da página. Além disso, é possível notar a grande quantidade de usuários cadastrados no site através dos comentários variados existentes, o que comprova a preferência dos usuários brasileiros em relação a fóruns sobre cinema. Acima de tudo, este site oferece boa variação opinativa de comentários, uma vez que comporta críticas positivas e negativas sobre os filmes,

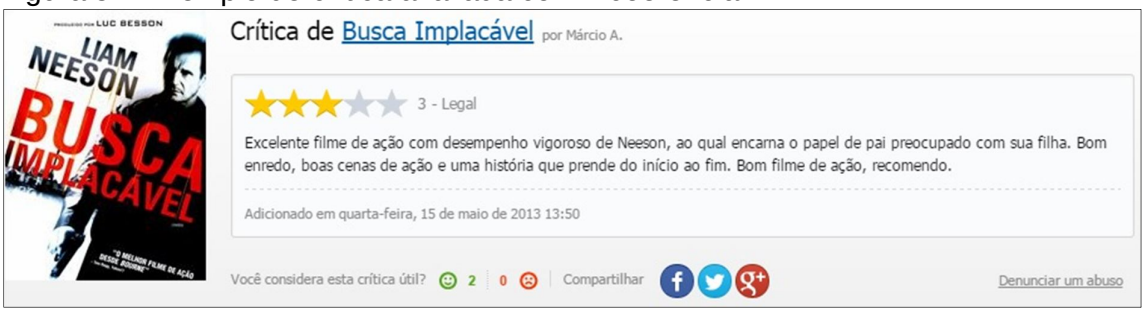
detalhe muito importante para a construção de um corpus que possibilite comprovar a real eficiência do classificador de opinião para ambas as polaridades.

Embora o presente trabalho esteja enquadrado na área de Inteligência Artificial (IA) e, com isso, proponha métodos automáticos para o cumprimento de cada passo, a etapa inicial - que se trata da coleta dos comentários e montagem do corpus - foi realizada manualmente, devido às inúmeras divergências encontradas durante a análise de algumas críticas. A mais recorrente se refere à incoerência de alguns usuários ao informar a nota final de avaliação, que deveria estar em acordo com a crítica apresentada.

A escala avaliativa do site permite ao usuário informar uma nota que varia de 0 a 5 estrelas, na qual “0 estrelas” indica o nível de maior insatisfação e, por consequência, 5 estrelas define o grau máximo de satisfação do usuário pelo filme que está sendo criticado.

A Figura 5 mostra um claro exemplo de crítica avaliada com incoerência: nela, o usuário se mostrou plenamente satisfeito com o filme, o que caracterizaria uma nota entre 4 e 5 estrelas de avaliação final, porém, seja por descuido ou falta de atenção, o usuário avaliou a própria crítica com apenas 3 estrelas.

Figura 5 – Exemplo de crítica avaliada com incoerência.



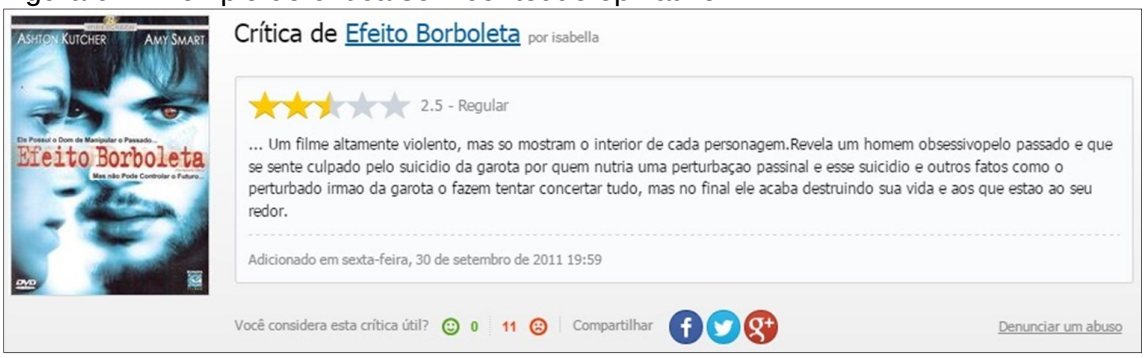
Fonte: Adorocinema.com.

Detalhes como este certamente ocultariam a real eficiência do classificador de opinião, pois uma crítica corretamente avaliada pelo sistema estaria em inconsistência com a avaliação real quando, na verdade, a informação equivocada está sendo transmitida pelo próprio autor da crítica.

Outro fator que desmotivou a coleta automática foi a grande concentração de comentários que não expressam conteúdo opinativo. Em vários destes casos, o autor inclusive narra a história ou apresenta a sinopse do filme, e conteúdos deste

tipo obrigariam o sistema a coletar um valor opinativo de onde não existe, comprometendo mais uma vez a eficiência do classificador. A Figura 6 exibe um comentário nestas condições, no qual a autora narra o decorrer do filme “Efeito Borboleta”.

Figura 6 – Exemplo de crítica sem conteúdo opinativo.



Fonte: Adorocinema.com.

Considerando que os comentários existentes na página Adorocinema.com foram escritos por internautas, e não por críticos especializados, a linguagem informal dos usuários se revelou como um pequeno obstáculo durante a análise de determinadas críticas, pois, em muitos casos, apresentaram erros de escrita, de pontuação, presença exagerada de gírias, falta de concordância, etc.

Considerando tais adversidades durante o processo manual de coleta, críticas aleatórias do site foram individualmente analisadas e, ao apresentarem perfil adequado, foram selecionadas para fazer parte do corpus de comentários. Os critérios adotados para considerar uma crítica como adequada abrangeram detalhes como: criatividade do autor em relação à crítica; vocabulário rico e diversificado; linguagem com um nível mínimo de qualidade e aproximação da norma culta da língua portuguesa; ausência de quantidade excessiva de gírias e de informações repetitivas e/ou redundantes; avaliação do autor acerca dos diversos quesitos que compõem um filme, como efeitos especiais, enredo, elenco, figurino, etc., não se limitando apenas a dizer que o filme é bom ou ruim.

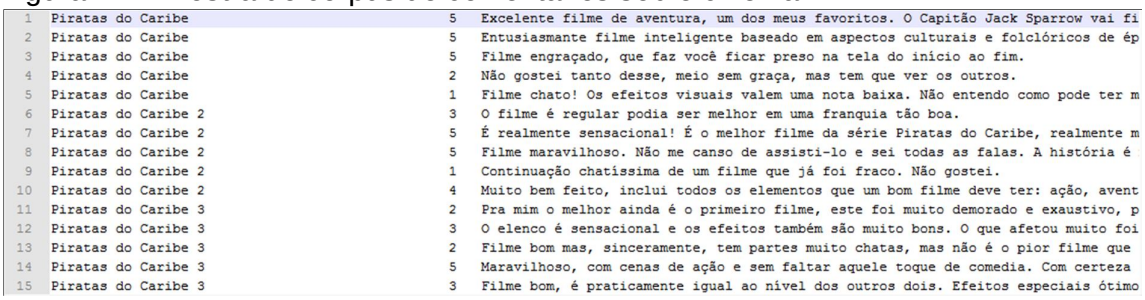
Outro critério importante para a seleção da crítica foi em relação ao seu tamanho, tomando o cuidado de não coletar comentários extremamente curtos, para evitar frases sem conteúdo opinativo.

As críticas escolhidas foram armazenadas uma por linha em um arquivo texto, separado em três colunas alinhadas: a primeira, com limite de tamanho 50, exibe o

nome do filme que está sendo criticado; logo em seguida, existe uma nota avaliativa para a crítica numa escala que varia de 1 a 5: “1” representa péssimo; “2” - ruim, “3” – mediano; “4” – bom; “5” – ótimo/excelente. As notas foram reaplicadas segundo a interpretação de cada comentário durante a sua análise.

Por fim, na terceira e última coluna do arquivo estão localizadas as críticas. No total, foram coletadas 400 críticas – divididas igualmente em 80 críticas para cada nota avaliativa. Uma amostra do corpus de comentários pode ser visualizada na Figura 7.

Figura 7 – Amostra do corpus de comentários sobre cinema.



1	Piratas do Caribe	5	Excelente filme de aventura, um dos meus favoritos. O Capitão Jack Sparrow vai fi
2	Piratas do Caribe	5	Entusiasmante filme inteligente baseado em aspectos culturais e folclóricos de ép
3	Piratas do Caribe	5	Filme engraçado, que faz você ficar preso na tela do início ao fim.
4	Piratas do Caribe	2	Não gostei tanto desse, meio sem graça, mas tem que ver os outros.
5	Piratas do Caribe	1	Filme chato! Os efeitos visuais valem uma nota baixa. Não entendo como pode ter m
6	Piratas do Caribe 2	3	O filme é regular podia ser melhor em uma franquia tão boa.
7	Piratas do Caribe 2	5	É realmente sensacional! É o melhor filme da série Piratas do Caribe, realmente m
8	Piratas do Caribe 2	5	Filme maravilhoso. Não me canso de assisti-lo e sei todas as falas. A história é
9	Piratas do Caribe 2	1	Continuação chatíssima de um filme que já foi fraco. Não gostei.
10	Piratas do Caribe 2	4	Muito bem feito, inclui todos os elementos que um bom filme deve ter: ação, avent
11	Piratas do Caribe 3	2	Pra mim o melhor ainda é o primeiro filme, este foi muito demorado e exaustivo, p
12	Piratas do Caribe 3	3	O elenco é sensacional e os efeitos também são muito bons. O que afetou muito foi
13	Piratas do Caribe 3	2	Filme bom mas, sinceramente, tem partes muito chatas, mas não é o pior filme que
14	Piratas do Caribe 3	5	Maravilhoso, com cenas de ação e sem faltar aquele toque de comedia. Com certeza
15	Piratas do Caribe 3	3	Filme bom, é praticamente igual ao nível dos outros dois. Efeitos especiais ótimo

Fonte: Elaborada pelo autor.

A fim de minimizar os erros ortográficos e assim melhorar a qualidade dos textos extraídos, eles foram momentaneamente inseridos no software Microsoft Excel, onde foi realizada uma revisão ortográfica e corrigidos os principais erros encontrados.

A partir deste ponto, o corpus já está pronto para ser processado e receber os primeiros tratamentos necessários para a classificação opinativa de cada crítica.

6.2. TRATAMENTO DO CORPUS

Levando em consideração a grande quantidade de sinais de pontuação presentes nos conteúdos textuais em geral, o primeiro tratamento relevante a ser aplicado ao corpus de críticas se refere justamente à exclusão destes símbolos, que poderão dificultar o processo de classificação opinativa nas etapas posteriores.

Os processos gerais realizados no presente trabalho foram executados através de comandos das linguagens AWK e SED, que se caracterizam pelo poder de manipulação de arquivos e textos. Estas linguagens combinam seus recursos

com scripts do Shell no Linux, aprimorando ainda mais as potencialidades de processamento desta plataforma sem utilizar muitas linhas de comando. O Shell é um módulo que permite ao usuário solicitar serviços do Linux através da execução de comandos internos. Em outras palavras, é um interpretador de linha de comando semelhante ao Prompt de Comando dos sistemas operacionais Windows.

O primeiro script utilizado é apresentado na Figura 8. Além de retirar os sinais de pontuação informados na própria Figura 8, esta sequência de comandos também serve para inserir o símbolo de sustenido “#” ao final de cada linha, pois este símbolo serve de referência para separar uma crítica da outra.

Figura 8 – Script para eliminar sinais de pontuação.

```
Retirar os seguintes sinais de pontuação:
" - ( ) , ; : ? ! .

sed -e 's/.\r/ #\r/g' Corpus > filmes; cp -a filmes TEMP;
sed -e 's/!\r/ #\r/g' TEMP > filmes; sed -e 's/"//g' filmes > TEMP;
sed -e 's/-//g' TEMP > filmes; sed -e 's/(//g' filmes > TEMP; sed -e 's/)//g' TEMP > filmes;
sed -e 's/,//g' filmes > TEMP; sed -e 's/;//g' TEMP > filmes; sed -e 's/:/g' filmes > TEMP;
sed -e 's/?//g' TEMP > filmes; sed -e 's/!//g' filmes > TEMP; sed -e 's/\.//g' TEMP > filmes
```

Fonte: Elaborada pelo autor.

Assim como muitas linguagens de programação, as linguagens AWK e SED utilizam o símbolo de “ponto e vírgula” (;) para separar seus comandos entre si. No primeiro comando da Figura 8, é utilizado como entrada o arquivo contendo o corpus de críticas, e o resultado do processamento é gravado em um arquivo chamado “filmes” que, por sua vez, é copiado e recebe o nome “TEMP” através do segundo comando. A partir disso, o arquivo de entrada/saída de cada processamento é alternado entre os arquivos “filmes” e “TEMP”.

Com o corpus livre de sinais de pontuação, a etapa de classificação gramatical das palavras já pode ser executada.

6.3. IDENTIFICAÇÃO DA CLASSE GRAMATICAL DAS PALAVRAS

De acordo com a atual concepção gramatical da língua portuguesa, qualquer palavra do idioma pode ser classificada em uma das dez classes gramaticais existentes: adjetivo, advérbio, artigo, conjunção, interjeição, numeral, preposição, pronome, substantivo e verbo. A classificação depende exclusivamente de

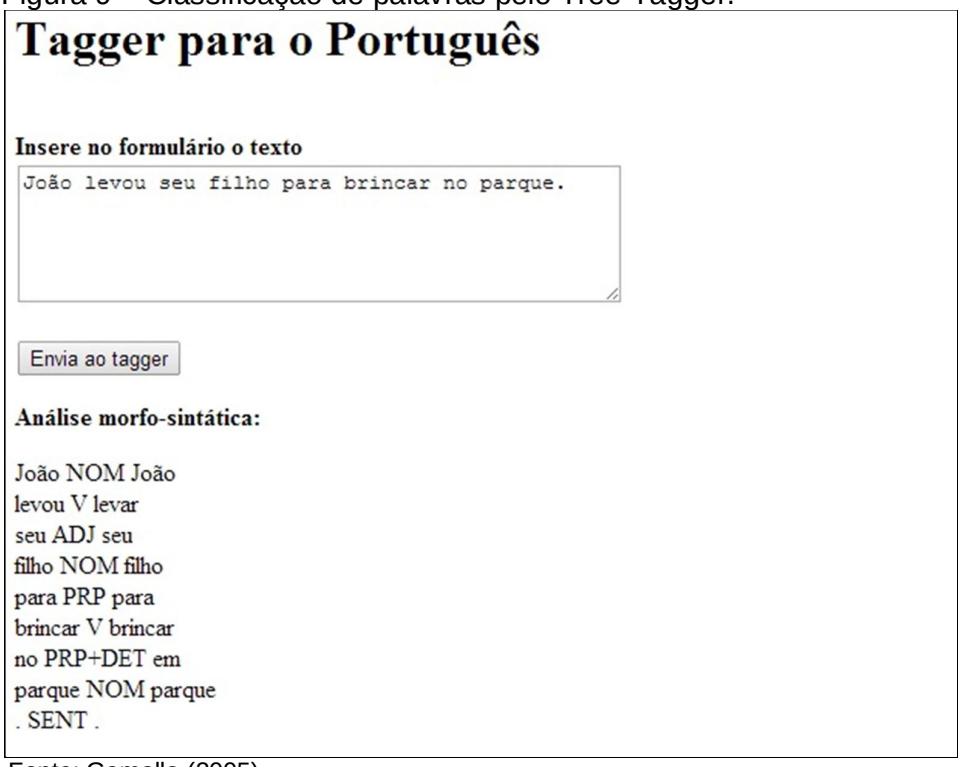
características da própria palavra, não envolvendo relação entre palavras. (ARAÚJO, 2006).

A área da gramática que estuda a classificação das palavras é a Morfologia, ou seja, o estudo da formação e derivação das palavras. (ARAÚJO, 2006). Este processo inicial da técnica do presente trabalho consiste em identificar a classe gramatical de cada palavra que compõe o comentário que está sendo analisado e, desta forma, cumpre a primeira etapa de um interpretador de língua natural convencional, que se trata da análise morfológica.

Para tanto, utiliza-se um software do tipo tagger (“etiquetador”). O sistema Tree-Tagger representa um aplicativo deste gênero e foi utilizado neste processamento. O sistema desenvolvido por Gamallo (2005) é executado por linhas de comando no Shell do Linux, mas possui uma versão web mais interativa, que será apresentada para melhor entendimento. A versão on-line possui uma interface simples e de fácil utilização (como pode ser visto na Figura 9), envolvendo objetos textuais, um campo para digitação e um botão responsável por iniciar o processamento do texto.

Para utilizar, basta ao usuário digitar a frase desejada e acionar o botão “Envia ao tagger”. Após alguns instantes, o sistema apresenta, logo abaixo do botão, a classificação precisa de cada token identificado na sentença. No exemplo mostrado na Figura 9, foi analisada a frase “João levou seu filho para brincar no parque”. É possível notar que o sistema, além de classificar cada palavra e sinal de pontuação, também mostra informações adicionais, como indicar que a palavra “levou” é uma flexão do verbo “levar”.

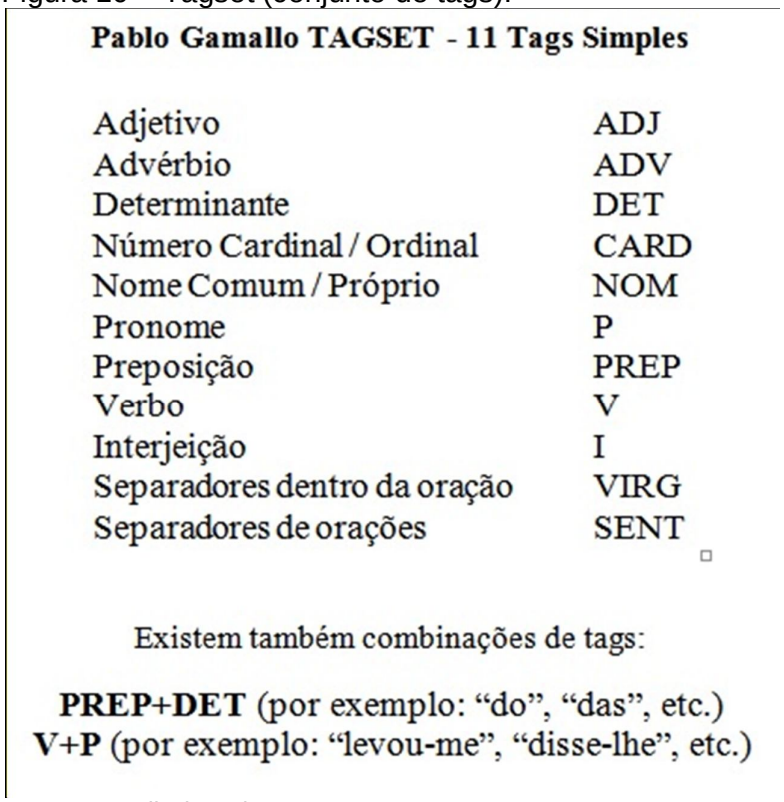
Figura 9 – Classificação de palavras pelo Tree-Tagger.



Fonte: Gamallo (2005).

Para melhor interpretação das siglas utilizadas pelo Tree-Tagger, na Figura 10 é apresentada a lista de tags, nomeada pelo criador (GAMALLO, 2005) como tagset (conjunto de tags). No caso, as tags representam as classes gramaticais das palavras.

Figura 10 – Tagset (conjunto de tags).



Pablo Gamallo TAGSET - 11 Tags Simples

Adjetivo	ADJ
Advérbio	ADV
Determinante	DET
Número Cardinal / Ordinal	CARD
Nome Comum / Próprio	NOM
Pronome	P
Preposição	PREP
Verbo	V
Interjeição	I
Separadores dentro da oração	VIRG
Separadores de orações	SENT

Existem também combinações de tags:

PREP+DET (por exemplo: “do”, “das”, etc.)
V+P (por exemplo: “levou-me”, “disse-lhe”, etc.)

Fonte: Gamallo (2005).

Para executar o Tree-Tagger pelo Shell do Linux, basta informar o caminho onde o aplicativo foi instalado e selecionar o arquivo referente ao idioma desejado, conforme pode ser visto na Figura 11.

Figura 11 – Comando para executar o Tree-Tagger no Shell do Linux.

```
-> Processar verificador Tree-Tagger:
cat filmes | ~/tree-tagger/cmd/tree-tagger-portuguese > TEMP
```

Fonte: Elaborada pelo autor.

O arquivo “filmes”, que foi a saída da sequência anterior de comandos, foi utilizado como entrada neste comando, e o resultado do seu processamento foi gravado no arquivo “TEMP”, cujo conteúdo pode ser visualizado na Figura 12.

Figura 12 – Resultado do processamento feito pelo Tree-Tagger.

```
Excelente   ADJ excelente
filme   NOM filme
de  PRP de
aventura    NOM aventura
um  DET um
dos PRP+DET de
meus    ADJ meu
favoritos   NOM favorito
O   V   <unknown>
Capitão NOM capitão
Jack    NOM jack
Sparrow NOM <unknown>
vai V   ir
ficar   V   ficar
marcado V   marcar
na  PRP+DET em
história    NOM história
do  PRP+DET de
cinema  NOM cinema
com PRP com
certeza NOM certeza
```

Fonte: Elaborada pelo autor.

Analisando o arquivo de saída gerado pelo Tree-Tagger, é possível perceber que, embora o aplicativo possua certa preocupação em colunar as palavras, muitas delas fogem do padrão por serem longas ou curtas demais. A fim de otimizar o conteúdo do arquivo para as próximas etapas, é desejável que seja gerado um espaçamento simples entre a primeira e a segunda palavra de cada linha e, de preferência, que a terceira palavra seja eliminada, por ser irrelevante para a situação presente. O primeiro comando mostrado na Figura 13 cumpre exatamente esta tarefa.

Ainda na Figura 13, existem outros dois trechos de comando que são responsáveis por corrigir falhas geradas pelo Tree-Tagger ao classificar pronomes possessivos (“meu”, “seu”, “teu” e derivados) e artigos (o, a, os, as, um, uma, uns, umas). Ao invés do aplicativo classifica-los, respectivamente, como pronome (P) e artigo (DET), a classe proposta erroneamente varia de acordo com a palavra, como pode ser visto mais adiante na Figura 14.

Figura 13 – Script para corrigir pronomes e artigos.

```
-> Retirar a 3ª coluna:
cat TEMP | awk '{print $1" "$2}' > filmes

-> Corrigir pronomes possessivos:
sed -e 's/\([M|m]eu \)\([A-Z]*\)/\1P/g' -e 's/\([M|m]inha \)\([A-Z]*\)/\1P/g'
-e 's/\([M|m]eus \)\([A-Z]*\)/\1P/g' -e 's/\([M|m]inhas \)\([A-Z]*\)/\1P/g'
-e 's/\([S|s]eu \)\([A-Z]*\)/\1P/g' -e 's/\([S|s]ua \)\([A-Z]*\)/\1P/g'
-e 's/\([S|s]eus \)\([A-Z]*\)/\1P/g' -e 's/\([S|s]uas \)\([A-Z]*\)/\1P/g' filmes > TEMP

-> Corrigir artigos:
sed -e 's/\(Um \)\([A-Z]*\)/\1DET/g' -e 's/\(Uma \)\([A-Z]*\)/\1DET/g'
-e 's/\(Uns \)\([A-Z]*\)/\1DET/g' -e 's/\(Umas \)\([A-Z]*\)/\1DET/g'
-e 's/\(O \)\([A-Z]*\)/\1DET/g' -e 's/\(A \)\([A-Z]*\)/\1DET/g'
-e 's/\(Os \)\([A-Z]*\)/\1DET/g' -e 's/\(As \)\([A-Z]*\)/\1DET/g' TEMP > filmes
```

Fonte: Elaborada pelo autor.

Figura 14 – Resultado após aplicação do script

Gerado pelo Tree-Tagger	*Corrigido com script*
Um ADJ	Um DET
Uma ADV	Uma DET
Uns DET	Uns DET
Umas CARD	Umas DET
O NOM	O DET
A P	A DET
Os PRP	Os DET
As V	As DET
Meu ADJ	Meu P
Minha NOM	Minha P
Meus NOM	Meus P
Minhas NOM	Minhas P
Seu NOM	Seu P
Sua NOM	Sua P
Seus NOM	Seus P
Suas NOM	Suas P

Fonte: Elaborada pelo autor.

Com o conhecimento da classe gramatical de cada palavra presente em um comentário, torna-se possível definir quais os termos relevantes para a análise semântica do comentário. Executa-se o próximo passo seguindo algumas regras específicas relacionadas justamente à sequência de palavras de acordo com sua classe gramatical.

6.4. EXTRAÇÃO DE BIGRAMAS RELEVANTES PARA O CÁLCULO DA ORIENTAÇÃO SEMÂNTICA

O procedimento geral de mineração de opinião dos comentários aqui adotado se baseia numa etapa principal que se encarrega de extrair bigramas (grupos de duas palavras) do texto, que serão selecionados de acordo com um conjunto de padrões morfossintáticos pré-definidos.

Dos bigramas que foram extraídos seguindo tais regras, pode-se considerar que a maioria realmente transmite conteúdo opinativo do comentário, enquanto outros podem não expressar nenhuma opinião.

Outro fator importante desta etapa se refere à desconsideração de algumas classes de palavras durante a análise dos bigramas que, deste modo, serão ignoradas para que seja analisada a classe da palavra seguinte. É o caso das classes: pronome, artigo, conjunção, preposição, interjeição e numeral, por não transmitirem valores semânticos em suas representações. Na área de mineração de opinião, esta categoria de palavra pode ser descartada do texto sem prejuízo de sentido. Estas palavras são conhecidas pelo termo “stopwords”, em razão de sua inutilidade para a classificação da opinião.

A Figura 15 apresenta o comando Shell do Linux na linguagem SED responsável por apagar linhas que possuem stopwords. É importante lembrar que o script anterior manipulou o corpus de modo que cada palavra se encontra em uma linha, juntamente com sua classe gramatical correspondente.

Figura 15 – Comandos para apagar stopwords.

```
-> Apagar linhas que tiverem stopwords
(DET, CARD, P, PRP, I, PRP+DET, V+P, CONJ, CONJSUB):

sed -i '/ DET/d' filmes; sed -i '/ CARD/d' filmes;
sed -i '/ P/d' filmes; sed -i '/ PRP/d' filmes;
sed -i '/ I/d' filmes; sed -i '/ PRP+DET/d' filmes;
sed -i '/ V+P/d' filmes; sed -i '/ CONJ/d' filmes;
sed -i '/ CONJSUB/d' filmes
```

Fonte: Elaborada pelo autor.

Os padrões morfossintáticos, apresentados na Figura 15, que se baseiam nas construções propostas por Turney (2002) para o idioma inglês, foram adaptados para o espanhol por Cruz et al. (2008). Devido à grande semelhança estrutural linguística entre os idiomas português e espanhol, neste presente trabalho serão consideradas as mesmas construções empregadas no artigo espanhol mencionado, porém com a adição de duas novas construções (1ª e 6ª linha da Figura 16).

Como se pode notar, as regras ou padrões morfossintáticos são guiados por sequências específicas de classes de palavras para a obtenção dos bigramas relevantes.

Figura 16 – Padrões Morfossintáticos para a Coleta de Bigramas Relevantes.

	Primeira Palavra	**Segunda Palavra**
1.	Adjetivo	Adjetivo
2.	Adjetivo	Substantivo
3.	Advérbio	Adjetivo
4.	Advérbio	Verbo
5.	Substantivo	Adjetivo
6.	Verbo	Adjetivo
7.	Verbo	Advérbio

Fonte: Cruz et al. (2008). Adaptada pelo autor.

Para identificar os bigramas relevantes presentes no corpus, é de fundamental importância a execução de alguns tratamentos para a correta manipulação do texto, que permitam posteriormente isolar cada bigrama relevante em uma linha e eliminar as demais combinações de palavras.

O primeiro tratamento desta etapa consiste em acumular todo o conteúdo textual do corpus na primeira linha do arquivo, a fim de prepara-lo para a execução de um script que insira uma quebra de linha para cada sustenido (#) encontrado (lembrando que os sustenidos haviam sido inseridos no corpus justamente para separar uma crítica da outra). A sequência dos comandos mencionados é apresentada na Figura 17.

Figura 17 – Script para colocar tudo na primeira linha.

```
-> Colocar todas as palavras na mesma linha:
sed ':a;N;$!ba;s/\n/ /g' filmes > TEMP

-> Separar palavras por crítica:
sed -e 's/# V /#\n/g' TEMP > filmes;
sed -e 's/# V/#/g' filmes > TEMP;
sed -e 's/# NOM /#\n/g' TEMP > AUX
```

Fonte: Elaborada pelo autor.

Após este processamento, o corpus de críticas já começa a adquirir aparência e organização próximas ao esperado final, quando os bigramas deverão se dispor um em cada linha. Como se pode notar na Figura 18, cada crítica já está disposta em sua respectiva linha, e o último símbolo é o sustenido (#).

Figura 18 – Aparência do corpus após a aplicação dos últimos scripts.

```
1  Excelente ADJ filme NOM aventura NOM favoritos NOM Capitão NOM Jack NOM Sparrow NOM vai
2  Entusiasmante NOM filme NOM inteligente ADJ baseado V aspectos NOM culturais ADJ folclór
3  Filme NOM engraçado ADJ faz V ficar V preso V tela NOM início NOM fim NOM #
4  Não NOM gostei V tanto ADV meio NOM graça NOM tem V ver V #
5  Filme NOM chato ADJ efeitos NOM visuais ADJ valem V nota NOM baixa ADJ Não NOM entendo V
6  filme NOM é V regular ADJ podia V ser V melhor ADJ franquia NOM tão ADV boa ADJ #
7  É V realmente ADV sensacional ADJ É V melhor ADJ filme NOM série NOM Piratas NOM Caribe
8  Filme NOM maravilhoso ADJ Não NOM canso V assistilo V sei V todas ADJ falas NOM história
9  Continuação NOM chatíssima ADJ filme NOM já ADV foi V fraco ADJ Não NOM gostei V #
10 Muito NOM bem ADV feito V inclui V todos ADJ elementos NOM bom ADJ filme NOM deve V ter
```

Fonte: Elaborada pelo autor.

Baseando-se nas regras/padrões morfossintáticos definidos e apresentados anteriormente, já é possível executar um script que inicie o processo de isolamento dos bigramas um por linha. A estratégia utilizada consiste em pular uma linha ao encontrar qualquer sequência que atenda a um dos critérios de classificação do bigrama como relevante. O script utilizado para esta finalidade é apresentado na Figura 19. Nele, é utilizado o comando “s///g” da linguagem SED, que funciona como o recurso de “localizar e substituir” presente em editores de texto em geral. Além disso, para a identificação dos padrões, são utilizadas expressões regulares.

Figura 19 – Script para pular linha ao encontrar um bigrama relevante.

```
-> Pular uma linha quando encontrar um bigrama relevante:
sed -e 's/\([a-Z]* ADJ [a-Z]* NOM\) /\1\n/g' AUX > filmes; rm AUX;
sed -e 's/\([a-Z]* ADV [a-Z]* ADJ\) /\1\n/g' filmes > TEMP;
sed -e 's/\([a-Z]* ADV [a-Z]* V\) /\1\n/g' TEMP > filmes;
sed -e 's/\([a-Z]* NOM [a-Z]* ADJ\) /\1\n/g' filmes > TEMP;
sed -e 's/\([a-Z]* V [a-Z]* ADV\) /\1\n/g' TEMP > filmes;
sed -e 's/\([a-Z]* V [a-Z]* ADJ\) /\1\n/g' filmes > TEMP;
sed -e 's/\([a-Z]* ADJ [a-Z]* ADJ\) /\1\n/g' TEMP > filmes
```

Fonte: Elaborada pelo autor.

Após a execução do último script, no arquivo de saída já é possível identificar como “bigramas relevantes” as duas últimas palavras de cada linha, separadas entre si pelo rótulo de sua respectiva classe gramatical (Exemplo de linha: [...] bom ADJ

filme NOM). A partir disso, uma nova sequência de comandos pode ser executada para, por fim, manter no arquivo apenas os bigramas relevantes (um em cada linha) e os sustenidos, que serão utilizados para separar os bigramas pertencentes a cada crítica. A Figura 20 demonstra exatamente esta lista de comandos.

Figura 20 – Lista de comandos para manter apenas os bigramas relevantes.

```
-> Deixar sustenido (#) isolado na linha:
sed -e 's/\( [a-Z]*\) #/\1\n#/g' filmes > TEMP

-> Apagar linhas que tiverem menos de 4 palavras:
cat TEMP | awk '{if ($0 ~ "#") {print} else if ($4) {print} }' > filmes

-> Manter apenas 4 últimas colunas:
cat filmes | awk '{if ($0 ~ "#") {print} else {print $(NF-3)" "$(NF-2)" "$(NF-1)" "$NF} }' > TEMP

-> Apagar linhas que não tiverem bigramas relevantes:
sed -i '/[a-Z]* ADJ [a-Z]* ADV/d' TEMP; sed -i '/[a-Z]* ADJ [a-Z]* V/d' TEMP;
sed -i '/[a-Z]* ADV [a-Z]* ADV/d' TEMP; sed -i '/[a-Z]* ADV [a-Z]* NOM/d' TEMP;
sed -i '/[a-Z]* NOM [a-Z]* ADV/d' TEMP; sed -i '/[a-Z]* NOM [a-Z]* NOM/d' TEMP;
sed -i '/[a-Z]* NOM [a-Z]* V/d' TEMP; sed -i '/[a-Z]* V [a-Z]* NOM/d' TEMP;
sed -i '/[a-Z]* V [a-Z]* V/d' TEMP

-> Manter apenas 1ª e 2ª colunas, mantendo os sustenidos (#):
cat TEMP | awk '{if ($0 ~ "#") {print} else {print $1" "$3} }' > filmes
```

Fonte: Elaborada pelo autor.

O resultado final deste processamento é exibido na Figura 21.

Figura 21 – Arquivo de bigramas relevantes.

```
1  Excelente filme
2  cinematográfico preferido
3  #
4  filme inteligente
5  aspectos culturais
6  detalhadamente pensada
7  esmeramente trabalhados
8  excelente atividade
9  música fundo
10 bem explicada
11 mágico ideia
12 #
13 Filme engraçado
14 gostei tanto
15 #
16 Filme chato
17 nota baixa
18 ter mais
19 #
20 é regular
```

Fonte: Elaborada pelo autor.

Os bigramas relevantes são processados na intenção de se calcular a orientação semântica proveniente de cada um, através de um método estatístico usado pelo algoritmo PMI-IR – que, em inglês, é a abreviação de “Pointwise Mutual Information” e “Information Retrieval” –, que consiste em estimar a Informação Mútua Pontual entre dois termos. A ideia deste algoritmo é encontrar um valor que indique a estimativa de possível aparição de uma palavra ou expressão em textos ou documentos da web a partir da existência de outra palavra. (CRUZ et al., 2008).

Para a aplicação deste algoritmo, exige-se o auxílio de um buscador da web para analisar a ocorrência dos dois termos em um mesmo documento. No cálculo da orientação semântica, a estimativa é realizada entre o bigrama e duas palavras-semente, que representam termos com indiscutível orientação semântica, sendo uma positiva e a outra negativa.

A técnica que será empregada se diferencia neste ponto por não envolver apenas uma palavra-semente de cada polaridade, mas dois conjuntos de três palavras-semente, sendo um deles formado por palavras positivas e o outro por palavras negativas.

6.5. DEFINIÇÃO DOS CONJUNTOS DE PALAVRAS-SEMENTE

Um experimento semelhante realizado por Cruz et al. (2008) abrangeu testes com uma palavra-semente de cada tipo e também com conjuntos dessas palavras, e os resultados obtidos revelaram uma melhora considerável na precisão de classificação dos comentários quando se utilizou os conjuntos. Por este motivo, o atual projeto considerou o uso de dois conjuntos com três palavras-semente cada um.

As palavras-semente escolhidas são apresentadas na Figura 22.

Figura 22 – Conjuntos de Palavras-Semente.

Palavras Positivas	**Palavras Negativas**
Bom	Mal
Ótimo	Ruim
Excelente	Péssimo

Fonte: Elaborada pelo autor.

As três palavras-semente de cada tipo foram selecionadas levando em conta a ordem de intensidade semântica contida em seus significados. Por exemplo, pode-

se interpretar que, das palavras positivas da tabela, “Bom” indicaria o grau positivo relativamente mais baixo, visto que “Ótimo” a superaria em sentido de satisfação, ocupando o nível mediano da tabela. Logo, o grau de satisfação máxima estaria associado à palavra “Excelente”.

Do mesmo modo, o conjunto de palavras negativas também segue este princípio, tratando, ainda, de representar o antônimo mais aproximado possível (em questão de sentido) em relação à palavra positiva ao seu lado no quadro.

Com todas estas informações, o algoritmo PMI-IR pode ser executado para, finalmente, classificar os comentários sobre cinema em positivo, negativo ou neutro.

6.6. CÁLCULO DA ORIENTAÇÃO SEMÂNTICA

Os cálculos iniciais para obtenção da orientação semântica de um termo se concentram em capturar uma informação estatística em relação ao aparecimento deste termo a partir da presença de outro termo - que, no caso, é uma palavra-semente -, considerando um mesmo documento ou página da web. Este processo pode ser facilmente realizado por um buscador de páginas da web, como o Bing.

O cálculo desta medida estatística é determinado pelo algoritmo PMI-IR, que apresenta sua fórmula geral na Figura 23.

Figura 23 – Fórmula do PMI.

$$PMI(w_1, w_2) = log_2 \left(\frac{p(w_1 \& w_2)}{p(w_1)p(w_2)} \right)$$

Fonte: Cruz et al. (2008).

Para calcular a medida estimativa PMI entre dois termos (que na fórmula estão identificados como w_1 e w_2), utiliza-se um buscador da web para retornar as ocorrências de páginas que contenham os dois termos próximo um do outro no texto (numerador da fórmula: $w_1 \& w_2$). Desta forma, o bigrama poderia ser representado pelo elemento w_1, e o elemento w_2 seria uma das palavras-semente propostas.

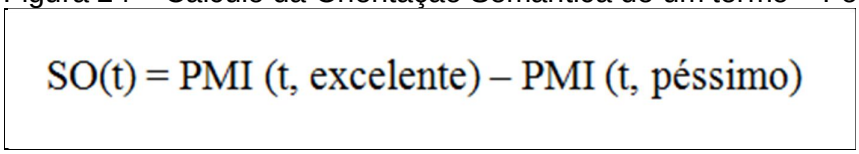
Neste caso, será utilizado o operador NEAR do Bing, onde é informado um valor que representa a quantidade máxima de palavras que podem intermediar os dois termos. Este valor foi adotado como o mais apropriado por evitar que sejam

consideradas ocorrências onde os termos estejam longe demais um do outro no texto de maneira que não formem uma ligação de sentido.

No denominador desta fórmula, determina-se encontrar também as ocorrências de páginas que comportam cada um dos dois termos, independente um do outro. Os valores são divididos e do resultado é calculado o logaritmo na base 2.

Com a obtenção desta medida estimativa entre os termos, torna-se possível calcular a orientação semântica (SO(t)) do bigrama, conforme mostra a Figura 24, onde o elemento *t* simboliza o termo analisado:

Figura 24 – Cálculo da Orientação Semântica de um termo – Forma Geral



Fonte: Cruz et al. (2008). Adaptada.

A fórmula apresentada se trata de uma representação genérica do cálculo de orientação semântica, onde se considera o uso de apenas uma palavra-semente para cada polaridade – positiva e negativa.

Para aplicá-la ao caso presente, que considera conjuntos de palavras-semente, calcula-se a medida estimativa entre o bigrama e cada uma das seis palavras-semente propostas. Enquanto os resultados obtidos das palavras-semente positivas devem manter o seu sinal aritmético para o cálculo, os resultados gerados pelas sementes negativas devem ter o seu sinal aritmético invertido e, em seguida, os valores devem ser todos somados até se obter o valor final, que representará a orientação semântica do bigrama.

A ideia central destes cálculos de orientação semântica é considerar que palavras ou expressões que transmitem opinião positiva apareçam frequentemente mais próximas a outras palavras com sentido positivo, como bom, ótimo ou excelente, e, em contrapartida, mais distantes de palavras que expressam opinião negativa.

Retomando o último arquivo gerado por comando em AWK no Shell do Linux, isto é, aquele que armazena apenas os bigramas relevantes separados por sustenido (#) entre uma crítica e outra, pode-se construir uma lista de URLs de pesquisa ao buscador Bing através da execução de um novo comando na linguagem AWK. As URLs obtidas servem para realizar buscas ao Bing utilizando os bigramas

individualmente e também estes em combinação com cada uma das palavras-semente. Uma amostra da lista de URLs obtida pode ser vista na Figura 25. A sintaxe das pesquisas realizadas segue a seguinte estrutura – “<bigrama>” near:10 <palavra-semente> – ou, substituindo por termos existentes no corpus – “filme inteligente” near:10 bom. Desta forma, o buscador Bing irá retornar páginas da web que contenham obrigatoriamente o bigrama e a palavra-semente a uma distância máxima de dez palavras intermediárias.

Figura 25 – Lista de URLs.

```
1  www.bing.com/search?q="cinematográfico+preferido"+near:10+bom
2  www.bing.com/search?q="cinematográfico+preferido"+near:10+otimo
3  www.bing.com/search?q="cinematográfico+preferido"+near:10+excelente
4  www.bing.com/search?q="cinematográfico+preferido"+near:10+mal
5  www.bing.com/search?q="cinematográfico+preferido"+near:10+ruim
6  www.bing.com/search?q="cinematográfico+preferido"+near:10+pessimo
7  www.bing.com/search?q="filme+inteligente"+near:10+bom
8  www.bing.com/search?q="filme+inteligente"+near:10+otimo
9  www.bing.com/search?q="filme+inteligente"+near:10+excelente
10 www.bing.com/search?q="filme+inteligente"+near:10+mal
```

Fonte: Elaborada pelo autor.

Através de um recurso existente num navegador de Internet em modo texto chamado Lynx, compatível com a plataforma Linux, é possível executar um comando no terminal Shell do Linux para gravar as informações textuais de determinada página da web em um arquivo texto. Executando este comando com um parâmetro do próprio navegador Lynx, pode-se programar a leitura das URLs presentes no arquivo texto gerado na etapa anterior. A Figura 26 destaca a localização do número de resultados de pesquisa no buscador Bing, e a Figura 27 exibe o comando que coleta este valor da busca.

Figura 26 – Destaque do número de resultados em uma pesquisa Bing.

Fonte: Elaborada pelo autor.

Figura 27 – Comando para retornar a quantidade de ocorrências de busca.

```
cat TEMP | lynx -dump - | awk '{if ($0 ~ /[0-9]+ results/) {print $1}
else if ($0 ~ /No results found for /) {print "0"} }' > filmes
```

Fonte: Elaborada pelo autor.

O arquivo de saída gerado pelo comando apresentado acima exibe os valores linha por linha, conforme mostra a Figura 28. Tal característica servirá para relacionar a qual bigrama o resultado se refere.

Figura 28 – Arquivo com as ocorrências de busca no Bing.

```
75  10
76  10
77  6
78  3
79  0
80  1.320
81  216
82  303
83  483
84  1.320
85  83
```

Fonte: Elaborada pelo autor.

Embora a Figura 28 não esteja exibindo, um símbolo de sustenido se mantem como separador entre os resultados de cada crítica, a fim de facilitar a identificação dos limites de uma crítica para outra.

Com a execução de um novo comando em SED, são retirados os sinais de ponto final presentes nos valores acima de 999, para que os valores sejam corretamente interpretados como números e assim possam ser utilizados no cálculo do algoritmo PMI-IR.

Considerando que cada grupo de sete linhas representa os resultados de um mesmo bigrama (resultado da pesquisa do próprio bigrama e deste com cada uma das seis palavras-semente), surge a necessidade de separá-los, de preferência, linha por linha, para facilitar a manipulação dos resultados e cálculo do algoritmo PMI. Para tanto, é executado um novo comando em SED para mover todo o conteúdo do arquivo para uma única linha, com os valores separados por espaço.

Com isso, ativa-se o primeiro script apresentado na Figura 29, que permite inserir uma quebra de linha a cada sequência de sete números lidos. Em seguida, outro script é executado para isolar os sustenidos em linha. O resultado desta operação pode ser visualizado na Figura 30.

Figura 29 – Inserir quebra de linha a cada sete números lidos.

```
-> Pular uma linha a cada 7 números encontrados:
sed -e 's/\([0-9]* [0-9]* [0-9]* [0-9]* [0-9]* [0-9]* [0-9]*\) /\1\n/g' filmes > TEMP


-> Pular linha ao encontrar sustenido (#):
sed 's/# /#\n/g' TEMP > filmes
```

Fonte: Elaborada pelo autor.

Figura 30 – Amostra do arquivo de saída.

```
147000 35 15 22 24 9 2
2850 13 3 5 8 2 1
#
139000 17 6 3 7 2 1
#
174000 262 34 36 40 33 10
#
23600 17 9 8 7 8 4
61800 49 11 20 9 12 4
2970000 12400 8650 2510 2200 6640 35
#
```

Fonte: Elaborada pelo autor.

Agora, cada linha apresenta os resultados de pesquisa de um bigrama isolado (primeiro campo da linha) seguido dos resultados de pesquisa entre o bigrama e cada uma das seis palavras-semente.

Estando as informações do arquivo dispostas desta maneira, torna-se muito simples realizar o cálculo do algoritmo PMI que, conforme mostrado anteriormente, consiste em dividir a quantidade de ocorrências de um bigrama próximo a uma palavra-semente (numerador da fórmula) pela multiplicação das ocorrências do bigrama com as da palavra-semente, ambos isoladamente.

Portanto, é aplicado o comando mostrado na Figura 31 para calcular esta medida estatística. Como pode ser visto na Figura, as colunas do arquivo são utilizadas como referência, e o número de ocorrências de cada palavra-semente isolada é definido previamente através de uma pesquisa manual ao Bing. O resultado deste processamento é um arquivo com seis colunas, que representam cada resultado de cálculo do PMI.

Figura 31 – Script para cálculo do algoritmo PMI-IR.

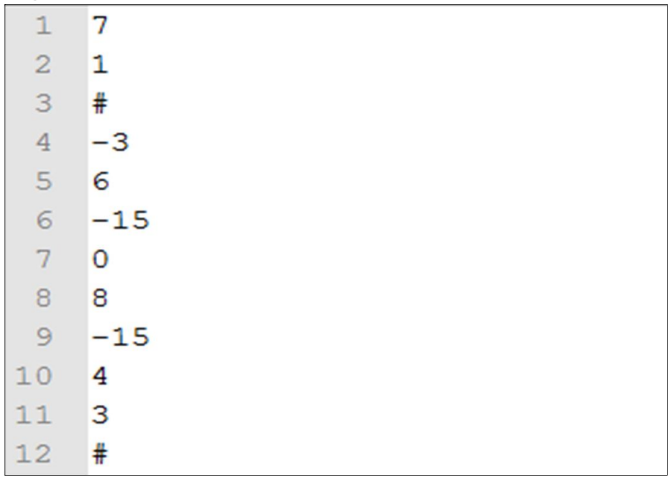
```
-> Cálculo do algoritmo PMI:
cat TEMP | awk '{if ($0 ~ "#") {print} else {print ($2/($1*23900000))" "($3/($1*8580000))
" "($4/($1*13600000))" "($5/($1*26600000))" "($6/($1*12800000))" "($7/($1*4770000))} }' filmes
```

Fonte: Elaborada pelo autor.

De acordo com a fórmula do PMI, dos resultados obtidos deve-se calcular o logaritmo na base 2. Por falta de recursos matemáticos nas linguagens de prompt utilizadas (AWK e SED), o cálculo de logaritmo precisou ser cumprido em um ambiente externo, o que demandou trabalho manual bem nas etapas finais do processo de classificação das críticas. No caso, foi utilizado o software de planilhas Microsoft Excel, onde cada coluna do arquivo foi processada. Em seguida, os dados foram novamente recompostos em um arquivo texto colunado.

Para obter a orientação semântica (SO) de cada bigrama, as três primeiras colunas de cada linha (que representam os resultados das palavras-semente positivas) são somadas entre si e, do resultado, subtrai-se os três valores seguintes (referentes às palavras negativas), conforme fórmula da orientação semântica – SO(t) – apresentada anteriormente. O resultado é um arquivo como o que está exibido na Figura 32.

Figura 32 – Resultados de Orientação Semântica para cada bigrama.

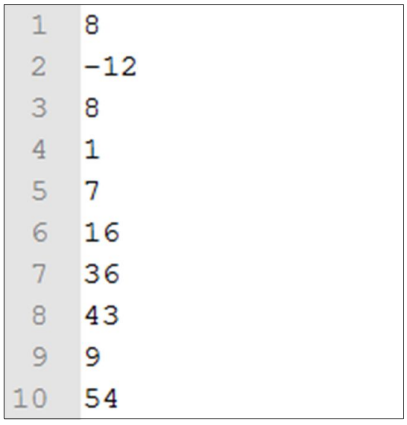


Fonte: Elaborada pelo autor.

Agora, só resta calcular os valores de cada bloco de crítica (separados entre si pelo símbolo de sustenido – “#”), para finalmente obter o valor semântico das críticas.

Para tanto, novamente é executado o comando de transferir todo o conteúdo do arquivo para uma única linha. Feito isso, através de um comando SED de “localizar e substituir”, localizam-se os sustenidos e, em substituição, insere-se uma quebra de linha, para separar os valores de cada crítica linha por linha. Por fim, realiza-se um somatório dos valores de cada linha entre si. O resultado deste processamento é a orientação semântica final de cada crítica, dispostas uma por linha, conforme Figura 33.

Figura 33 – Arquivo final com a orientação semântica de cada crítica.



Fonte: Elaborada pelo autor.

6.7. CLASSIFICAÇÃO OPINATIVA DO COMENTÁRIO

Baseando-se nos valores obtidos, que representam o nível semântico de cada comentário, o próximo passo é compará-los às notas de avaliação fornecidas pelo autor de cada crítica (neste trabalho em específico, as notas foram reavaliadas para se aplicar uma nota mais adequada).

Em geral, sistemas de Mineração de Opinião adotam que valores semânticos inferiores a zero representam opinião negativa, enquanto valores maiores que zero são associados a opiniões positivas.

Neste trabalho, foram consideradas três classes de saída para o algoritmo, que pode classificar as opiniões como positiva, negativa ou neutra. Além disso, foi levada em consideração a intensidade ou grau do sentimento, para diferenciar opiniões que separem, por exemplo, autores plenamente satisfeitos de autores que estejam ligeiramente empolgados com o tema.

Para tanto, foram analisadas as devidas fronteiras que separam cada intensidade de opinião. As opiniões neutras, de igual modo, foram assim consideradas quando o seu valor semântico se aproximou do limite entre crítica positiva e negativa. Todos os limites foram estudados com base em observações e experimentos manuais realizados com os comentários do corpus montado para este trabalho.

7 RESULTADOS

A fim de relatar o nível de precisão do presente classificador de opinião, na Figura 34 são apresentados alguns bigramas que foram selecionados pelo sistema como relevantes para a classificação da crítica em questão. Ao lado de cada bigrama, está o seu respectivo valor semântico (SO) calculado.

Figura 34 – Valor semântico obtido para alguns bigramas coletados do corpus.

Nº	Bigrama	SO
1.	Excelente filme	7
2.	Bem explicada	4
3.	Filme chato	1
4.	É regular	2
5.	É melhor	10
6.	Chatíssimo filme	2
7.	Existem piores	1
8.	É sensacional	21
9.	Afetou muito	3
10.	Direção péssima	-18
11.	Momentos emocionantes	15
12.	Importância pequena	5
13.	Perceber pequeno	-7
14.	Enredo legal	4
15.	Prometia tanto	-4
16.	Esperava mais	-6

Fonte: Elaborada pelo autor.

Como se pode notar na Figura 34, seguindo o modelo padrão dos sistemas de Mineração de Opinião para classificar a polaridade das palavras, a maioria dos bigramas apresentados possui inconsistência em relação ao seu valor semântico real. O bigrama “Filme chato”, por exemplo, foi avaliado pelo sistema com a nota “1” que, seguindo o método padrão de avaliação, por ser um valor inteiro positivo (maior que zero), caracteriza uma opinião positiva, sentimento contrário ao que este bigrama realmente transmite.

Levando em consideração esta e outras observações, como as obtidas durante a análise do valor semântico de críticas completas, definiram-se limites específicos para os diferentes níveis de sentimento.

Quando se considerou as críticas neutras nos experimentos, estas foram assim classificadas quando apresentaram valor semântico (obtido pelo sistema) entre 4 e 6, incluindo os dois valores. Para valores semânticos iguais ou inferiores a -15, a crítica foi considerada como péssima (nota 1); para valores entre -15 e 4, ruim (nota 2); acima de 6 e inferior/igual a 18, a crítica foi avaliada como boa (nota 4); e, por fim, valores superiores a 18 foram classificados como crítica ótima (nota 5).

Uma primeira comparação, abrangendo as notas avaliativas reais de cada crítica e o valor semântico obtido pelo sistema para classifica-los numa escala de 1 a 5, revelou baixa precisão alcançada pelo classificador (em torno de 25% para cada nota), fato que demonstrou a grande dificuldade de se classificar corretamente as críticas de acordo com seu grau de intensidade, ao invés de determinar apenas se a crítica é positiva, negativa ou neutra. A Figura 35 exibe os resultados mencionados.

Figura 35 – Precisão do classificador ao avaliar numa escala de 1 a 5.

Avaliação	**Críticas Classificadas Corret. / Total de Críticas**	**Precisão (%)**
Nota 1 (Péssimo)	19/80	23,75%
Nota 2 (Ruim)	22/80	27,50%
Nota 3 (Neutro)	04/80	5%
Nota 4 (Bom)	19/80	23,75%
Nota 5 (Ótimo)	20/80	25%
Total	84/400	21%

Fonte: Elaborada pelo autor.

Além do baixo desempenho geral obtido, o resultado que mais chama a atenção é em relação à precisão obtida ao classificar as críticas neutras (nota 3), que não passou dos 5%. Conforme já se previa, este dado comprovou a grande dificuldade de se classificar corretamente críticas neutras por métodos automáticos, provavelmente por tais críticas apresentarem, em geral, variação intensa de sentimento/opinião, o que acaba por influenciar o classificador a assumir a polaridade (positiva ou negativa) de maior tendência semântica da crítica, ainda que por pouca diferença.

Desconsiderando a nota avaliativa 3, que representa o grau de nível neutro, obteve-se uma leve melhora de precisão para as notas 2 e 4, que acabaram por classificar corretamente 23 críticas num total de 80 cada uma, conforme mostra a Figura 36. Os resultados das notas 1 e 5 não foram influenciados neste novo teste

pois simplesmente expandiu-se os limites das notas 2 e 4 (que fazem fronteira com a nota 3) da seguinte forma: para valores semânticos acima de -15 mas igual/inferior a 4, crítica ruim (nota 2); acima de 4 e igual/inferior a 18, crítica avaliada como boa (nota 4).

Figura 36 – Precisão do classificador ao avaliar sem a pontuação neutra.

Avaliação	**Críticas Classificadas Corret. / Total de Críticas**	**Precisão (%)**
Nota 1 (Péssimo)	19/80	23,75%
Nota 2 (Ruim)	23/80	27,50%
Nota 4 (Bom)	23/80	23,75%
Nota 5 (Ótimo)	20/80	25%
Total	85/320	26,5%

Fonte: Elaborada pelo autor.

Agora, adotando como resultados de saída as classificações “crítica positiva”, “crítica neutra” e “crítica negativa”, é obtida uma melhora considerável na precisão do classificador, uma vez que a situação não obriga mais o sistema a classificar corretamente a intensidade do sentimento expresso na crítica, mas apenas indicar a sua tendência semântica. É possível notar que neste novo experimento o classificador mostrou o mesmo resultado dos dois experimentos anteriores em relação às críticas neutras, isso porque foi considerado o mesmo intervalo de valores para a classificação neutra (valores semânticos entre 4 e 6, inclusive).

A Figura 37 apresenta justamente os resultados baseados nestas três classes de saída.

Figura 37 – Precisão do classificador ao considerar três classes.

Avaliação	**Críticas Classificadas Corret. / Total de Críticas**	**Precisão (%)**
Negativa	80/160	50%
Neutra	4/80	5%
Positiva	98/160	61,25%
Total	182/400	45,5%

Fonte: Elaborada pelo autor.

Neste experimento, foram consideradas como “crítica negativa” as pontuações semânticas abaixo de 4, e foram testadas as críticas do corpus que

possuem notas 1 e 2 (ruim e péssimo). Da mesma forma, valores semânticos superiores a 6 foram considerados pelo classificador como “crítica positiva”, sendo que nesta verificação foram usadas as críticas do corpus que possuem notas 4 e 5 (bom e ótimo). O resultado final apontado por este experimento indicou uma precisão total de 45,5%, demonstrando uma melhoria expressiva em relação aos experimentos anteriores.

Das 80 críticas negativas classificadas corretamente, 42 haviam recebido nota 1 no corpus, e 38 pertencem à nota 2. Das 98 críticas positivas, 53 pertencem à nota 4, e 45 pertencem à nota 5.

No último experimento realizado, as classificações foram restritas a apenas dois possíveis resultados de saída, sendo eles “crítica positiva” e “crítica negativa”. O resultado obtido neste teste superou o do experimento anterior, ao apresentar uma precisão total de 58,75%, conforme mostra a Figura 38.

Figura 38 – Precisão do classificador ao considerar duas classes.

Avaliação	**Críticas Classificadas Corret. / Total de Críticas**	**Precisão (%)**
Negativa	84/160	52,5%
Positiva	104/160	65%
Total	188/320	58,75%

Fonte: Elaborada pelo autor.

Seguindo os padrões dos experimentos anteriores, foi considerada como “crítica negativa” a que apresentou valor semântico igual ou inferior a 4 entre as que receberam pontuações 1 e 2 no corpus. Da mesma forma, foi considerada como “crítica positiva” a que demonstrou orientação semântica superior a 4 entre as que receberam pontuações 4 e 5 no corpus.

8 CONSIDERAÇÕES FINAIS

Através dos experimentos realizados, pode-se considerar que a técnica de Mineração de Opinião empregada no presente trabalho não se trata de um método plenamente satisfatório, uma vez que o único resultado a superar precisão de classificação para metade das críticas foi o último, quando se considerou apenas duas classes de saída para o classificador.

Apesar disso, um fator que contribuiu principalmente para os resultados de menor precisão foi considerar a avaliação de valor “neutro” para o corpus de críticas. Isso porque, conforme já foi observado na seção anterior, as críticas consideradas como neutras possuem grande variação de sentimento no comentário, então, ao invés do classificador realizar um balanceamento dos valores semânticos encontrados na crítica, o sistema acaba por assumir uma das polaridades.

De qualquer forma, a presente técnica poderia ser testada em um corpus com maior quantidade de críticas, de preferência considerando como classes avaliativas apenas os valores “positivo” e “negativo”, a fim de abranger maior variedade de críticas e testar o classificador com um conteúdo mais diversificado.

É importante ressaltar que a presente metodologia, embora possa ser adaptada e melhorada a fim de classificar comentários para fins práticos, se trata de uma técnica de análise superficial de textos, não podendo identificar elementos subjetivos da linguagem como, por exemplo, ironia ou sarcasmo.

Uma possível linha de trabalho futuro poderia inclusive envolver aspectos subjetivos e pragmáticos da linguagem, como a significância prática e objetiva das expressões, que são interpretadas não somente através da análise das palavras como elementos individuais, mas que também requerem uma análise geral e contextual do texto. Esta certamente representaria uma técnica mais aprimorada e complexa de classificação opinativa, que vale ser explorada a fim de se buscar resultados mais satisfatórios e reais.

REFERÊNCIAS

AIRES, R. V. X. Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil. **Yumpu**, 2000. Disponível em: <https://www.yumpu.com/pt/document/view/14109298/view-researchgate/41>. Acesso em: 16 mai. 2014.

CRUZ, F.L. et al. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. **Sociedad Española para el Procesamiento Del Lenguaje Natural**, c2008. Disponível em: http://www.sepln.org/revistaSEPLN/revista/41/sec3-art2.pdf. Acesso em: 7 mar. 2014.

ARAÚJO, A. P. Classes de Palavras. **Info Escola**, c2006-2014. Disponível em: <http://www.infoescola.com/portugues/classes-de-palavras/>. Acesso em: 26 mai. 2014.

ASSIS, P. O que é tag? **Tecmundo**, c2009. Disponível em: <http://www.tecmundo.com.br/navegador/2051-o-que-e-tag-.htm>. Acesso em: 16 mai. 2014.

DICAS e atalhos turbinam suas buscas com o Bing. **UOL Tecnologia**, c1996-2014. Disponível em: <http://tecnologia.uol.com.br/album/2013/02/19/conheca-dicas-para-utilizar-bem-o-bing-para-pesquisas-na-internet.htm>. Acesso em: 15 dez. 2014.

GAMALLO, P. **Tree-Tagger**, 2005. Disponível em: <http://gramatica.usc.es/~gamallo/tagger.htm>. Acesso em: 16 mai. 2014.

GUEDES, Rafael; AFONSO, Derkian; MAGALHÃES, Lúcia Helena de. Mineração de opiniões de usuários na busca de conhecimento. **Vianna Sapiens**, 2010. Disponível em: <http://www.viannajunior.edu.br/files/uploads/20131001_141137.pdf>. Acesso em: 14 nov. 2014.

GUERBER, C. R. Compiladores e Interpretadores. **Universidade do Contestado**, c2007. Disponível em:<http://www.mfa.unc.br/info/carlosrafael/aco/aula16.pdf>. Acesso em: 8 abr. 2014.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques, Second Edition**. Morgan Kaufmann, 2006.

HEFREN, A. Determinismo e Gramática Sintagmática. **Vernáculo da Física**, c2010. Disponível em: <http://alexandrehefren.wordpress.com/2010/03/14/determinismo-e-gramatica-sintagmatica-gs-parte-1/>. Acesso em: 15 mai. 2014.

MEDEIROS, J. C. D. Processamento Morfológico e Correção Ortográfica do Português. **Linguateca**, c1995. Disponível em: <http://www.linguateca.pt/Repositorio/TeseMestradoJoseCarlosMedeiros.pdf>. Acesso em: 8 abr. 2014.

PETRY, A. Reconhecimento automático de voz. **Consultoria Faccin**, c2008. Disponível em: <http://www.consultoriafaccin.com.br/artigos-interessantes/reconhecimento-automatico-voz.html> Acesso em: 8 abr. 2014.

ROSA, Maria Carlota. **Introdução à Morfologia**. Editora Contexto. São Paulo, 2005.

RUSSEL, S. J.; NORVIG, P. **Inteligência Artificial**. Rio de Janeiro: Elsevier, 2004.

SANTOS, L. M. Protótipo para mineração de opinião em redes sociais. **Universidade Federal de Lavras**, 2010. Disponível em: <http://www.bcc.ufla.br/wp-content/uploads/2013/2010/LeandroMatioli.pdf>. Acesso em: 21 mai. 2014.

SATO, P. O que é Inteligência Artificial? **Revista Escola**, 2009. Disponível em: <http://revistaescola.abril.com.br/ciencias/fundamentos/inteligencia-artificial-onde-ela-aplicada-476528.shtml> Acesso em: 4 abr. 2014.

SILVA, B. C. D. et al. **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. Série de Relatórios Técnicos do NILC, NILC-TR-07-10. São Carlos, 2007.

TURNEY, Peter D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. **Association for Computational Linguistics**, 2002. Disponível em: <http://aclweb.org/anthology//P/P02/P02-1053.pdf>. Acesso em: 22 abr. 2014.

Mineração de opinião aplicada para a classificação de críticas em português sobre cinema

Vinicius Rios Gianezi, Patrick Pedreira Silva, Elvio Gilberto da Silva, André Luiz Ferraz Castro

Universidade Sagrado Coração (USC) – Bauru, SP – Brasil

vinicius.gianezi@gmail.com, {patrick.silva, egsilva}@usc.br, andcastro@ig.com.br

Abstract. *In the recent world stage, marked by globalization and the constant competitivity in the companies, the great amount of data generated daily on the web includes an important source opinion about varied topics, including review products and services availables in the market. In this context, arise the need for automating the collection process and text classification opinionated generated by web users, because these informations reveal the quality of the product or service provided. This paper presents the application of an opinion mining technique to classify comments in portuguese about cinema in positive, negative or neutral, based on sentiment expressed in the text. Moreover, are presented the general methodology concepts of this study, as well as the search limitations and the possible fields for future study.*

Resumo. *No atual cenário mundial, marcado pela globalização e pelo aumento constante da competitividade nas empresas, a grandiosa massa de dados gerada diariamente na web comporta uma rica fonte de opiniões de usuários acerca dos mais variados assuntos, que inclui avaliações de produtos e serviços disponíveis no mercado. Neste contexto, surge a necessidade de automatização do processo de coleta e classificação dos textos opinativos gerados por estes usuários, uma vez que tais informações refletem a qualidade do produto ou serviço prestado. O presente trabalho demonstra a aplicação de uma técnica de mineração de opinião que consiste em classificar críticas sobre cinema em positivas, negativas ou neutras, de acordo com o sentimento empregado. Além disso, são apresentados os conceitos gerais da metodologia que rege este trabalho, bem como as limitações da pesquisa e os possíveis campos para trabalho futuro.*

1. Introdução

No mercado competitivo dos últimos anos, as empresas passaram a buscar estratégias para se destacarem, sendo que a principal delas está associada à satisfação do usuário final. A internet, por ser a mais rica fonte de comentários e opiniões de consumidores, tem sido o principal alvo de análise das grandes empresas.

Com isso, a opinião dos internautas se tornou objeto de muito interesse por parte das empresas, afinal o comentário do consumidor representa um termômetro que indica o quanto o produto e/ou serviço está agradando e em quais pontos precisa melhorar.

Para automatizar o processo de classificação de textos opinativos, muito se tem explorado acerca do processamento de linguagem natural, um dos vastos campos de estudo da Inteligência Artificial. O histórico de seus experimentos aponta que sua aplicação inicial revela certo grau de complexidade, mas retorna resultados satisfatórios quando se utilizam métodos apropriados. (SILVA et al., 2007; CRUZ et al., 2008).

O objetivo do atual trabalho é processar comentários em português sobre cinema com a intenção de coletar o sentimento empregado pelo autor, medindo sua carga de satisfação e insatisfação para, assim, classificá-lo como neutro, positivo ou negativo, através da aplicação de uma técnica linguística em um corpus montado a partir da coleta de comentários da web.

2. Processamento de Linguagem Natural (PLN)

O surgimento dos computadores, além de permitir avanços consideráveis nos diversos campos do estudo científico, também abriu caminho para novas frentes de pesquisa que, até então, não eram sequer cogitadas. Entre elas, podem-se destacar os estudos sobre PLN, que vêm se desenvolvendo conforme ocorre o aperfeiçoamento da comunicação entre homem e máquina. (SILVA et al., 2007).

O primeiro grande desafio era fazer a máquina compreender instruções para a execução de tarefas. Foi então que surgiram as primeiras linguagens de programação, que permitiram a comunicação homem-máquina através da confusa linguagem de máquina. (SILVA et al., 2007).

Hoje em dia, os projetos mais ambiciosos desta área passaram a ser a criação de sistemas capazes de interpretar e gerar mensagens codificadas em línguas naturais, que tornem possível uma interação verbal do homem com a máquina. (SILVA et al., 2007).

2.1. Interpretador de Língua Natural

Um interpretador realiza diferentes níveis de processamento linguístico de forma a coletar descrições específicas das palavras em cada etapa. Cada uma destas informações é armazenada no léxico, que "[...] consiste em um conjunto de palavras ou expressões da língua associadas a um conjunto de atributos [...]". (SILVA et al., 2007, p. 33). Portanto, o léxico se trata de um fundamental recurso para o PLN, uma vez que as fases de interpretação e geração do código em língua natural dependem do acesso e manipulação de suas informações.

Os processos seguintes de interpretação envolvem análises sintática, semântica, discursiva e pragmática, e não são, necessariamente, executados nesta sequência, podendo haver processos combinados de forma distinta dependendo da especificação do projeto que está sendo desenvolvido. (SILVA et al., 2007).

2.2. Tagger

O processo de identificar a classe gramatical das palavras de uma sentença pode ser executado por um tipo de aplicativo conhecido como Tagger. O nome deste tipo de aplicativo origina-se do conceito de tag, que representa um termo associado a determinado conteúdo (como uma imagem, documento ou música), que serve para identificar o seu gênero contextual facilitando, assim, a busca de conteúdos por tema. Em outras palavras, sua função se relaciona com a tradução da palavra vinda do inglês – "etiqueta". Portanto, estas palavras-chave permitem uma maior organização das informações na Internet de maneira que informações relacionadas sejam agrupadas. (ASSIS, 2009).

No contexto presente, o sistema tagger se encarrega de detectar a classe gramatical de cada palavra em uma sentença, como adjetivos, pronomes, substantivos, entre outras classificações.

Com o conhecimento da classe gramatical de cada palavra que compõe a sentença que, no caso, será um comentário sobre cinema, será possível coletar os bigramas relevantes para a detecção da opinião expressa no comentário. As regras que definem a relevância dos bigramas

serão apresentadas na seção Metodologia deste artigo. Cada bigrama que se encaixar nas condições impostas será processado com a finalidade de se calcular a sua orientação semântica.

2.3. Orientação Semântica

A orientação semântica de uma palavra ou expressão consiste em definir se ela possui uma conotação positiva ou negativa na frase em termos de sentido qualitativo. (CRUZ et al., 2008). Por exemplo, o termo "ótimo" referencia algo bom, atribuindo um sentido positivo ao objeto abordado na frase. Enquanto isso, a expressão "mal" indica um valor negativo para algo que certamente não agradou o autor do comentário.

No seu conceito absoluto, a orientação semântica se trata de um valor matemático, da classe dos reais, que representa a medida subjetiva do sentido de uma palavra ou expressão. Assim, possuindo um valor positivo (maior que zero), possui implicações positivas e, da mesma forma, se apresentar valor negativo (menor que zero), indica um sentimento negativo. (CRUZ et al., 2008). O buscador Bing oferece um operador, denominado NEAR, que pode ser utilizado por sistemas de PLN no cálculo da orientação semântica. Tal operador será apresentado no tópico seguinte.

2.4. Operador NEAR

O buscador Bing, assim como vários outros sistemas de busca da web, possui operadores especiais que podem ser utilizados para restringir o efeito da busca.

O comando NEAR (traduzido para o português como "próximo"), que será utilizado no presente trabalho, se trata de um operador de proximidade que impõe um limite na busca de dois ou mais termos (palavras ou expressões) em um mesmo documento, baseando-se em uma quantidade máxima de palavras que podem intermediá-los. (DICAS..., c1996-2014). Desta forma, permite procurar por conteúdos que contenham os dois termos pesquisados sem estarem necessariamente ligados por alguma hierarquia ou dependência contextual, ao contrário do que ocorreria em uma busca comum.

A sintaxe do operador NEAR é apresentada na Figura 1. Após o símbolo de dois pontos, deve ser indicada a quantidade máxima de palavras que pode separar os dois termos. (DICAS..., c1996-2014).

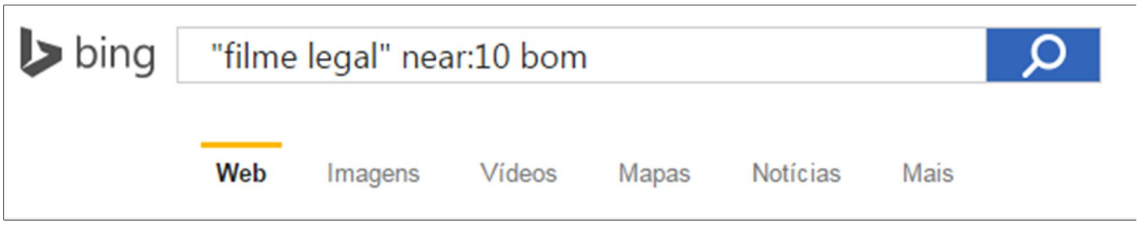


Figura 1 – Sintaxe do Operador NEAR.

A ordem das palavras-chave no comando não é um fator relevante para a consulta e, desta maneira, pesquisar no Bing "ciência near:10 computação" poderá trazer resultados com ordens diferentes de ocorrência dos termos, como "[...] ciência da computação [...]" e "A computação é a ciência que [...]". Ou seja, a ordem dos termos nos conteúdos encontrados não seguirá a ordem digitada no comando NEAR.

3. Mineração de Dados

Em meio ao emaranhado de informações dos mais variados tipos e assuntos presentes na web, surge a necessidade de filtrar os dados na intenção de se extrair conteúdos relevantes para o aperfeiçoamento de algum tipo de atividade.

A partir deste ponto, surge o conceito de mineração de dados, ou data mining, que representa a técnica de extração ou filtragem de dados úteis em meio a vastos volumes de informação. (HAN; KAMBER, 2006). Em outras palavras, designa a área de estudo da Inteligência Artificial responsável por obter conhecimento, desvendar estruturas e extrair padrões ocultos em massas de dados. (GUEDES; AFONSO; MAGALHÃES, 2010).

Ao analisar conjuntos de dados diversos, como aqueles dispostos na web, por exemplo, é possível buscar padrões implícitos de informação que seja relevante para as mais variadas aplicações comerciais, como para traçar o perfil dos consumidores de determinado produto, ampliar um plano de negócios, analisar riscos e otimizar técnicas de divulgação e propaganda.

3.1. Mineração de Opinião

A mineração de opinião, ou opinion mining, também conhecida como análise de sentimento, representa um dos ramos de estudo do PLN e uma derivação da mineração de dados (data mining), e se concentra basicamente em extrair o sentimento opinativo empregado em um texto que seja direcionado à crítica de algum objeto passível de qualificação. (SANTOS, 2010). O alvo pode ser um produto ou aparelho, a campanha política de um candidato à eleição ou até mesmo conteúdos culturais, como músicas, livros e filmes.

Os estudos desta área se baseiam em classificar como positivo ou negativo o sentimento expresso em um texto crítico, com a intenção de indicar se o autor assume uma posição a favor ou contra o objeto em discussão, de acordo com a polaridade de significado das palavras contidas no texto.

4. Metodologia

O presente trabalho de mineração de opinião se concentrou em identificar a opinião crítica de comentários em português sobre cinema, classificando-os como positivo, negativo ou neutro de acordo com a orientação semântica expressada no texto.

Para tanto, inicialmente foi montado um corpus de comentários em português sobre cinema a partir da coleta manual de críticas existentes no site Adorocinema.com.

4.1. Obtenção e Montagem do corpus de comentários

A técnica de mineração de opinião exposta neste trabalho foi aplicada em um corpus de comentários sobre cinema. Tais comentários foram extraídos manualmente do site AdoroCinema.com, que foi escolhido pela qualidade do seu conteúdo jornalístico, fato que revela a confiabilidade da página.

As críticas escolhidas foram armazenadas uma por linha em um arquivo texto, separado em três colunas alinhadas: a primeira, com limite de tamanho 50, exibe o nome do filme que está sendo criticado; logo em seguida, existe uma nota avaliativa para a crítica numa escala que varia de 1 a 5: “1” representa péssimo; “2” - ruim, “3” – mediano; “4” – bom; “5” – ótimo/excelente. As notas foram reaplicadas segundo a interpretação de cada comentário durante a sua análise.

4.2. Tratamento do Corpus

Os processos gerais realizados no presente trabalho foram executados através de comandos das linguagens AWK e SED, que se caracterizam pelo poder de manipulação de arquivos e textos. Estas linguagens combinam seus recursos com scripts do Shell no Linux, aprimorando ainda mais as potencialidades de processamento desta plataforma sem utilizar muitas linhas de comando.

Levando em consideração a grande quantidade de sinais de pontuação presentes nos conteúdos textuais em geral, o primeiro tratamento relevante que foi aplicado no corpus de críticas foi justamente a exclusão destes símbolos, que poderiam dificultar o processo de classificação opinativa nas etapas posteriores.

4.3. Identificação da classe gramatical das palavras

De acordo com a atual concepção gramatical da língua portuguesa, qualquer palavra do idioma pode ser classificada em uma das dez classes gramaticais existentes: adjetivo, advérbio, artigo, conjunção, interjeição, numeral, preposição, pronome, substantivo e verbo. A classificação depende exclusivamente de características da própria palavra, não envolvendo relação entre palavras. (ARAÚJO, 2006).

Para identificar a classe gramatical das palavras presentes no corpus, utilizou-se um software do tipo tagger (“etiquetador”). O sistema Tree-Tagger representa um aplicativo deste gênero e foi utilizado neste processamento. O sistema desenvolvido por Gamallo (2005) é executado por linhas de comando no Shell do Linux e foi combinado com outros comandos das linguagens AWK e SED.

4.4. Extração de bigramas relevantes para o cálculo da orientação semântica

Para considerar um bigrama como relevante, foram considerados alguns padrões morfossintáticos (apresentados na Figura 2) que se baseiam nas construções propostas por Turney (2002) para o idioma inglês, e que foram adaptados para o espanhol por Cruz et al. (2008). Devido à grande semelhança estrutural linguística entre os idiomas português e espanhol, neste presente trabalho foram consideradas as mesmas construções empregadas no artigo espanhol mencionado, porém com a adição de duas novas construções (1ª e 6ª linha da Figura 2). Como se pode notar, as regras ou padrões morfossintáticos são guiados por sequências específicas de classes de palavras para a obtenção dos bigramas relevantes.

	Primeira Palavra	**Segunda Palavra**
1.	Adjetivo	Adjetivo
2.	Adjetivo	Substantivo
3.	Advérbio	Adjetivo
4.	Advérbio	Verbo
5.	Substantivo	Adjetivo
6.	Verbo	Adjetivo
7.	Verbo	Advérbio

Figura 2 – Padrões Morfossintáticos para a Coleta de Bigramas Relevantes.

4.5. Definição dos conjuntos de palavras-semente

Um experimento semelhante realizado por Cruz et al. (2008) abrangeu testes com uma palavra-semente de cada tipo e também com conjuntos dessas palavras, e os resultados obtidos revelaram uma melhora considerável na precisão de classificação dos comentários

quando se utilizou os conjuntos. Por este motivo, o atual projeto considerou o uso de dois conjuntos com três palavras-semente cada um. As palavras-semente escolhidas são apresentadas na Figura 3.

Palavras Positivas	Palavras Negativas
Bom	Mal
Ótimo	Ruim
Excelente	Péssimo

Figura 3 – Conjuntos de Palavras-Semente.

4.6. Cálculo da Orientação Semântica

Os cálculos iniciais para obtenção da orientação semântica de um termo se concentram em capturar uma informação estatística em relação ao aparecimento deste termo a partir da presença de outro termo - que, no caso, é uma palavra-semente -, considerando um mesmo documento ou página da web. Este processo pode ser facilmente realizado por um buscador de páginas da web, como o Bing.

O cálculo desta medida estatística é determinado pelo algoritmo PMI-IR, que apresenta sua fórmula geral na Figura 4.

$$PMI(w_1, w_2) = log_2 \left(\frac{p(w_1 \& w_2)}{p(w_1)p(w_2)} \right)$$

Figura 4 – Fórmula do PMI.

Com a obtenção desta medida estimativa entre os termos, torna-se possível calcular a orientação semântica (SO(t)) do bigrama, conforme mostra a Figura 5, onde o elemento *t* simboliza o termo analisado:

$$SO(t) = PMI\ (t, excelente) - PMI\ (t, péssimo)$$

Figura 5 – Cálculo da Orientação Semântica de um termo – Forma Geral

A fórmula apresentada se trata de uma representação genérica do cálculo de orientação semântica, onde se considera o uso de apenas uma palavra-semente para cada polaridade – positiva e negativa.

Para aplicá-la ao caso presente, que considera conjuntos de palavras-semente, calcula-se a medida estimativa entre o bigrama e cada uma das seis palavras-semente propostas. Enquanto os resultados obtidos das palavras-semente positivas devem manter o seu sinal aritmético para o cálculo, os resultados gerados pelas palavras-semente negativas devem ter o seu sinal aritmético invertido e, em seguida, os valores devem ser todos somados até se obter o valor final, que representará a orientação semântica do bigrama.

Somando-se, por fim, os valores de orientação semântica calculado para cada bigrama de um mesmo comentário, obtém-se a orientação semântica do comentário.

4.7. Classificação opinativa do comentário

Baseando-se nos valores obtidos, que representam o nível semântico de cada comentário, o próximo passo é compará-los às notas de avaliação fornecidas pelo autor de cada crítica (neste trabalho em específico, as notas foram reavaliadas para se aplicar uma nota mais adequada).

Para classificar as críticas do corpus, foram analisadas as devidas fronteiras que separam cada intensidade de opinião. As opiniões neutras, de igual modo, foram assim consideradas quando o seu valor semântico se aproximou do limite entre crítica positiva e negativa. Todos os limites foram estudados com base em observações e experimentos manuais realizados com os comentários do corpus montado para este trabalho.

5. RESULTADOS

No principal experimento realizado, as classificações foram restritas a apenas dois possíveis resultados de saída, sendo eles “crítica positiva” e “crítica negativa”. O resultado obtido neste teste atingiu uma precisão total de 58,75%, conforme mostra a Figura 6.

Avaliação	**Críticas Classificadas Corret. / Total de Críticas**	**Precisão (%)**
Negativa	84/160	52,5%
Positiva	104/160	65%
Total	188/320	58,75%

Figura 6 – Precisão do classificador ao considerar duas classes.

Seguindo os padrões dos demais experimentos, foi considerada como “crítica negativa” a que apresentou valor semântico igual ou inferior a 4 entre as que receberam pontuações 1 e 2 no corpus. Da mesma forma, foi considerada como “crítica positiva” a que demonstrou orientação semântica superior a 4 entre as que receberam pontuações 4 e 5 no corpus.

6. CONSIDERAÇÕES FINAIS

Através dos experimentos realizados, pode-se considerar que a técnica de Mineração de Opinião empregada no presente trabalho não se trata de um método plenamente satisfatório, uma vez que o resultado revela precisão de classificação para apenas pouco mais da metade das críticas.

Apesar disso, um fator que contribuiu principalmente para os resultados de menor precisão foi considerar a avaliação de valor “neutro” para o corpus de críticas. De qualquer forma, a presente técnica poderia ser testada em um corpus com maior quantidade de críticas, de preferência considerando como classes avaliativas apenas os valores “positivo” e “negativo”, a fim de abranger maior variedade de críticas e testar o classificador com um conteúdo mais diversificado.

É importante ressaltar que a presente metodologia, embora possa ser adaptada e melhorada a fim de classificar comentários para fins práticos, se trata de uma técnica de análise superficial de textos, não podendo identificar elementos subjetivos da linguagem como, por exemplo, ironia ou sarcasmo.

Uma possível linha de trabalho futuro poderia inclusive envolver aspectos subjetivos e pragmáticos da linguagem, como a significância prática e objetiva das expressões, que são interpretadas não somente através da análise das palavras como elementos individuais, mas que também requerem uma análise geral e contextual do texto. Esta certamente representaria uma técnica mais aprimorada e complexa de classificação opinativa, que vale ser explorada a fim de se buscar resultados mais satisfatórios e reais.

Referências

ARAÚJO, A. P. Classes de Palavras. **Info Escola**, c2006-2014. Disponível em: <http://www.infoescola.com/portugues/classes-de-palavras/>. Acesso em: 26 mai. 2014.

ASSIS, P. O que é tag? **Tecmundo**, c2009. Disponível em: <http://www.tecmundo.com.br/navegador/2051-o-que-e-tag-.htm>. Acesso em: 16 mai. 2014.

CRUZ, F.L. et al. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. **Sociedad Española para el Procesamiento Del Lenguaje Natural**, c2008. Disponível em: http://www.sepln.org/revistaSEPLN/revista/41/sec3-art2.pdf. Acesso em: 7 mar. 2014.

DICAS e atalhos turbinam suas buscas com o Bing. **UOL Tecnologia**, c1996-2014. Disponível em: <http://tecnologia.uol.com.br/album/2013/02/19/conheca-dicas-para-utilizar-bem-o-bing-para-pesquisas-na-internet.htm>. Acesso em: 15 dez. 2014.

GAMALLO, P. **Tree-Tagger**, 2005. Disponível em: <http://gramatica.usc.es/~gamallo/tagger.htm>. Acesso em: 16 mai. 2014.

GUEDES, Rafael; AFONSO, Derkian; MAGALHÃES, Lúcia Helena de. Mineração de opiniões de usuários na busca de conhecimento. **Vianna Sapiens**, 2010. Disponível em: <http://www.viannajunior.edu.br/files/uploads/20131001_141137.pdf>. Acesso em: 14 nov. 2014.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques, Second Edition**. Morgan Kaufmann, 2006.

SANTOS, L. M. Protótipo para mineração de opinião em redes sociais. **Universidade Federal de Lavras**, 2010. Disponível em: <http://www.bcc.ufla.br/wp-content/uploads/2013/2010/LeandroMatioli.pdf>. Acesso em: 21 mai. 2014.

SILVA, B. C. D. et al. **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. Série de Relatórios Técnicos do NILC, NILC-TR-07-10. São Carlos, 2007.

TURNEY, Peter D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. **Association for Computational Linguistics**, 2002. Disponível em: <http://aclweb.org/anthology//P/P02/P02-1053.pdf>. Acesso em: 22 abr. 2014.