VERIFICAÇÃO DE ACEITABILIDADE DO PRODUTO NO MERCADO POR MEIO DE TÉCNICAS DE *SENTIMENT ANALISYS* E *OPINION MAINING*

VERIFICATION OF PRODUCT ACCEPTABILITY OF THE MARKET BY TECHNICAL OF SENTIMENT ANALISYS AND OPINION MAINING

LUIZ HENRIQUE DIAS **AFONSO**[1*], FILIPE ROSEIRO **CÔGO**[2]

1. Graduado em Engenharia de controle e automação. Especialista em gestão de projetos, pela Unicesumar, docente da Faculdade de Engenharia e Inovação Técnico Profissional – FEITEP; **2.** Mestre em Ciência da Computação, docente da Universidade Federal do Paraná.

* Rua Ciprestes, 145, Maringá, Paraná, Brasil, CEP 87023-640. luizhenriqueafonso@gmail.com

Recebido em 05/09/2015. **Aceito** para publicação em 10/11/2015

RESUMO

Este trabalho tem como objetivo utilizar as técnicas da área de inteligência artificial e ciência da computação como *Sentiment analysis* e *Opinion mining* para a mineração de opiniões sobre marcas de cosméticos. Tais opiniões, em sua grande maioria, são postadas em blogs, fóruns e redes sociais na internet. O intuito é obter informações, analisá-las e interpretá-las, para que de maneira efetiva possam ser utilizadas a fim de melhorar a seleção do produto a ser disponibilizado em prateleiras ao consumidor, servir como apoio em futuras pesquisas e desenvolvimentos de novos produtos, de acordo com as opiniões manifestadas. Para validação de resultados foi utilizado o software desenvolvido na universidade de Waikato, Nova Zelândia, chamado Weka (Waikato Environment for Knowledge Analysis) com o objetivo de agregar algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas fazendo o processamento, análise, interpretação e contagem das palavras que mais foram encontradas em textos, resenhas e comentários em páginas, blogs e redes sociais da web. Desse modo, procurou-se identificar e analisar conceitos sobre marcas de cosméticos e seus produtos mais procurados no mercado pelo consumidor. Neste caso o público-alvo é o público feminino e o produto mais especificamente é o esmalte de unha.

PALAVRAS-CHAVE: Análise de sentimentos, Mineração de opinião.

ABSTRACT

This paper aims to use the techniques in the area of artificial intelligence and computer science as Sentiment analysis and data mining Opinion for mining of views on cosmetic brands. Such opinions, for the most part, are posted on blogs, forums and social networking sites. The aim is to obtain information, analyze them and interpret them so effectively can be used to improve the selection of the product being made available on shelves to consumers, serve as support for future research and development of new products, according to the views expressed. For results validation we used the software developed at the University of Waikato, New Zealand, called Weka (Waikato Environment for Knowledge Analysis) in order to add algorithms from different approaches / paradigms in the subfield of artificial intelligence dedicated to the part by learning study machines doing the processing, analysis, interpretation and count the words that were most commonly found in texts, reviews and comments on pages, blogs and social networks of the web. Thus, we tried to identify and analyze concepts of cosmetic brands and their most sought after products in the consumer market. In this case the target audience is the female audience and the product specifically is the nail polish.

KEYWORDS: Sentiment analysis, opinion mining.

1. INTRODUÇÃO

A competitividade entre empresas no ramo de cosméticos vem aumentando a cada dia, assim como o nível de exigência dos seus clientes. Empresas desse ramo, principalmente as de grande porte, têm investido fortemente em pesquisas para obterem melhorias e desenvolverem “o produto perfeito”, para assim conquistar o seu principal consumidor, o público feminino. A cada semana são lançados uma quantidade grande de novos produtos e para se obter uma vantagem competitiva e atrair mais consumidores é necessária uma seleção de quais produtos serão representados ou vendidos por comerciantes, de acordo com as opiniões e

gosto de seus consumidores.

Atualmente empresas buscam opinião de seus clientes através da pesquisa de campo, ou seja, coleta de dados referente aos mesmos e à análise e interpretação dos dados obtidos. O problema da pesquisa e que tal procedimento exige tempo e um custo elevado.

Os objetivos gerais desta pesquisa é agilizar a capitação das informações e reduzir tempo e custo dessa capitação. Os objetivos específicos são utilizar técnicas de inteligência artificial como *Sentiment analysis* e *Opinion maining* para fazer uma varredura online, filtrar e capitar palavras chaves para obter opiniões do público sobre produtos cosméticos, no caso desta pesquisa o esmalte.

Sentiment analysis ou análise de sentimento, a qual se refere ao processamento de linguagem natural, linguística computacional e análise de texto para identificar e extrair informações subjetivas em matérias primas [1].

Harrison (1998 apud Santos, 2008) [2] define *Data Mining* como um processo de exploração e análise de uma grande massa de dados fazendo uso de meios automáticos ou semi automáticos, objetivando a descoberta de padrões e regras significativas.

Com o aumento da popularidade da web e facilidade de acesso, se tornou mais comum usuários expressarem suas opiniões ou postar comentários sobre pessoas, organizações ou produtos em blogs, sites, redes sociais, entre outros, por exemplo, quando uma pessoa pretende comprar um laptop ou celular de uma determinada marca, ela visita páginas da web em busca de informações e opiniões daqueles que já compraram o produto a fim de realizarem uma avaliação e então decidir se irá comprar o produto ou não.

Desta forma, com o crescimento de informações na web, se torna cada vez mais necessário o uso de ferramentas eficientes e eficazes para mineração de conhecimentos úteis da web.

O presente artigo propõe uma análise de opinião baseada na frequência de palavras, ou seja, a análise do valor morfológico das palavras, dentre elas a detecção de adjetivos, que geralmente são palavras que por si só apresentam um caráter positivo ou negativo [3] para selecionar produtos que agradam mais o consumidor e assim representar e vender produtos cosméticos que mais agradão as pessoas que fazem uso do mesmo. Uma solução aparentemente simples, porém, muito eficiente comparado ao trabalho de ler grandes documentos dispersos na web.

Existem várias ferramentas para se utilizar na busca de informações, neste trabalho optou-se pela utilização do *google* por ser uma ferramenta popular e conceituado por pesquisadores. O Passo seguinte foi utilizar o software Weka para mineração de opiniões e exposição dos resultados.

SENTIMENT ANALYSIS

Sentiment analysis é um problema de categorização de texto no qual se deseja detectar opiniões favoráveis e desfavoráveis com relação a um determinado tópico [4].

Segundo Oguri (2006) [4], nos últimos anos o interesse da comunidade por *Sentiment analysis*, onde os documentos são classificados pelo sentimento, conotação e opinião, vem crescendo. A tarefa básica na análise de sentimento é classificar a polaridade de um texto se a opinião expressa é positiva, negativa, ou neutra, além da classificação do sentimento com, por exemplo, raiva, triste e feliz.

Uma aplicação de *Sentiment analysis* é na parte inicial do ciclo de vida de um produto, onde se deseja monitorar qual a vontade e a necessidade do consumidor, o que vem sendo publicado na web a respeito do lançamento de um determinado produto, quais as opiniões positivas e negativas do público sobre este produto ou até mesmo como a imagem da empresa vem evoluindo ao longo dos tempos [5].

Tradicionalmente, os setores de marketing realizam pesquisa de mercado com este propósito. Embora pesquisa de mercado seja uma área bastante madura e quando bem elaboradas gerem boas estimativas, tendem a ser muito custosas principalmente quando se trata de grandes volumes de dados. Com a explosão de informações disponíveis na Web num ambiente onde todos tendem a ser geradores de conteúdo e expressarem opiniões sobre os mais variados assuntos, aplicações que consolidam opiniões e geram estatísticas relevantes passam a ter um valor significativo [4].

B. Pang. *et al*. 2002 (apud OGURI, 2006) [4] Uma das principais abordagens aplicadas a análise de sentimento consiste em modelar documentos com palavras selecionadas e aplicar métodos estatísticos que utilizam a frequência das palavras para treinar classificadores.

OPINION MINING

Dado um conjunto de documentos de textos avaliativos que contem opiniões, sobre um objeto, a mineração de opinião tem como objetivo extrair atributos e componentes do objeto que foi comentado em cada documento para determinar se se os comentários são positivos negativos ou neutros. Informações textuais no mundo podem ser classificadas em duas categorias principais, fatos e opiniões. Os fatos são informações objetivas sobre as entidades e eventos em todo o mundo. Opiniões são informações subjetivas que refletem o sentimento das pessoas ou percepções sobre as entidades e eventos [6].

As técnicas de mineração de dados da Web surgiram das técnicas de data mining. A mineração de dados teve início nos anos 80 quando os profissionais da área começaram a se preocupar com o crescente armazenamento dos dados inutilizados nos computadores das empresas. Mineração de dados é um conceito que se assemelha ao termo inglês KDD (*Knowledge discovery in Database*), ou seja, descoberta de conhecimento nos bancos de dados [7].

A acirrada competitividade entre as empresas, faz com que passem a direcionar seus produtos e serviços a públicos específicos. A descoberta dos perfis destes clientes e suas preferencias pode ser obtida através do processo de mineração [2]. As principais tarefas de mineração são:

Classificação: É a tarefa de mineração mais usada por se parecer muito com a compreenção que o ser humano tem do ambiente no qual esta inserido [2].

Esta tarefa consiste em categorizar objetos de acordo com suas caracteristicas e agrupá-los em conjunto de classes prédefinidas, por este motivo pode ser considerada como preditiva. Ela faz uso de resultados discrétos e sua principal técnica é a árvore de classificação [8].

Estimativa / previsão: Com base nos valores de variáveis conhecidas, a estimativa tem a função de estipular valores de variáveis desconhecidas. Faz uso de valores contínuos e também é considerada uma tarefa preditiva já que existem informações definidas anteriormente que servirão de base para aplicação da tarefa. Por exemplo pode estimar as chances de chover avaliando um conjunto de diagnósticos como temperatura, umidade, se está ventando e condições do tempo como nublabo, ensolarado ou chuvoso [9].

Associação / Descrição: O processo de mineração pode também, ter a função de descrever a maneira como determinados dados foram produzidos a fim de aumentar o conhecimento das pessoas. Uma boa descrição gera uma boa explicação [9].

Segmentação: Esta tarefa consiste em agrupar elementos de uma população heterogênea em pequenos grupos mais homogêneos. Em contradição a tarefa de classificação, na segmentação não se faz necessário a existencia de classes prédefinidas. Nada precisa ser dito ao sistema, o próprio algoritimo descobre os objetos semelhantes e faz o agrupamento, sem ter a necessidade da interação do usuário. Como exemplo, pode-se citar a carteira de clientes de uma grande loja, onde analisando cada cliente, é possivel agrupa-los de acordo com seus hábitos de consumo e usar esta informação em uma outra tarefa de mineração [9].

2. MATERIAL E MÉTODOS

Neste artigo foi utilizado a pesquisa de campo, que procede a observação de fatos e fenômenos da mesma maneira com que ocorrem na realidade, a coleta de dados e análise e interpretação desses dados [10].

Existe também a determinação da coleta de dados mais apropriadas a natureza do tema [10]. Foram usados sites especializados em cosméticos, blogs especializados e redes sociais.

O presente artigo tem por objetivo demonstrar aplicações das técnicas de *Sentiment analysis* e *Opinion mining* com o objetivo de agilizar a pesquisa de mercado e selecionar os principais produtos de cosméticos a serem vendidos ou representados, para exemplificar foram escolhidos esmaltes de unha. Assim o trabalho concentrasse em 3 etapas; a busca de informações, mineração dos dados e validação dos resultados obtidos através da busca de informações e da mineração das opiniões.

Com o auxílio desta ferramenta foram lidas 125 resenhas sobre esmaltes em vários blogs da internet. Este número de resenhas lidas foi baseado na norma NBR 5426, planos de amostragem e procedimentos na inspeção por atributos.

Optou-se pela escolha do ramo de cosméticos, mais especificamente por esmaltes, pois a ramo de cosmético vem crescendo muito nos últimos anos e a cada semana novos produtos são lançados no mercado, tais produtos são avaliados por blogueiras e depois são atribuídas notas de 1 a 5 estrelas. Como a gama de produtos é muito grande, representantes e empresário que trabalham neste ramo tem tido muitas dificuldades em selecionar produtos para venderem que agradem seus clientes.

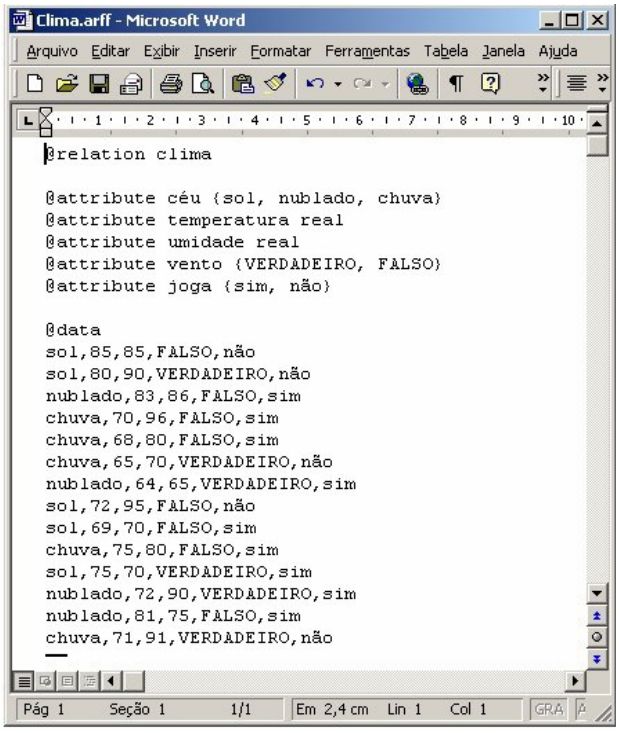


Figura 1. Arquivo no formato ARFF. **Fonte:** Dados do autor, 2015.

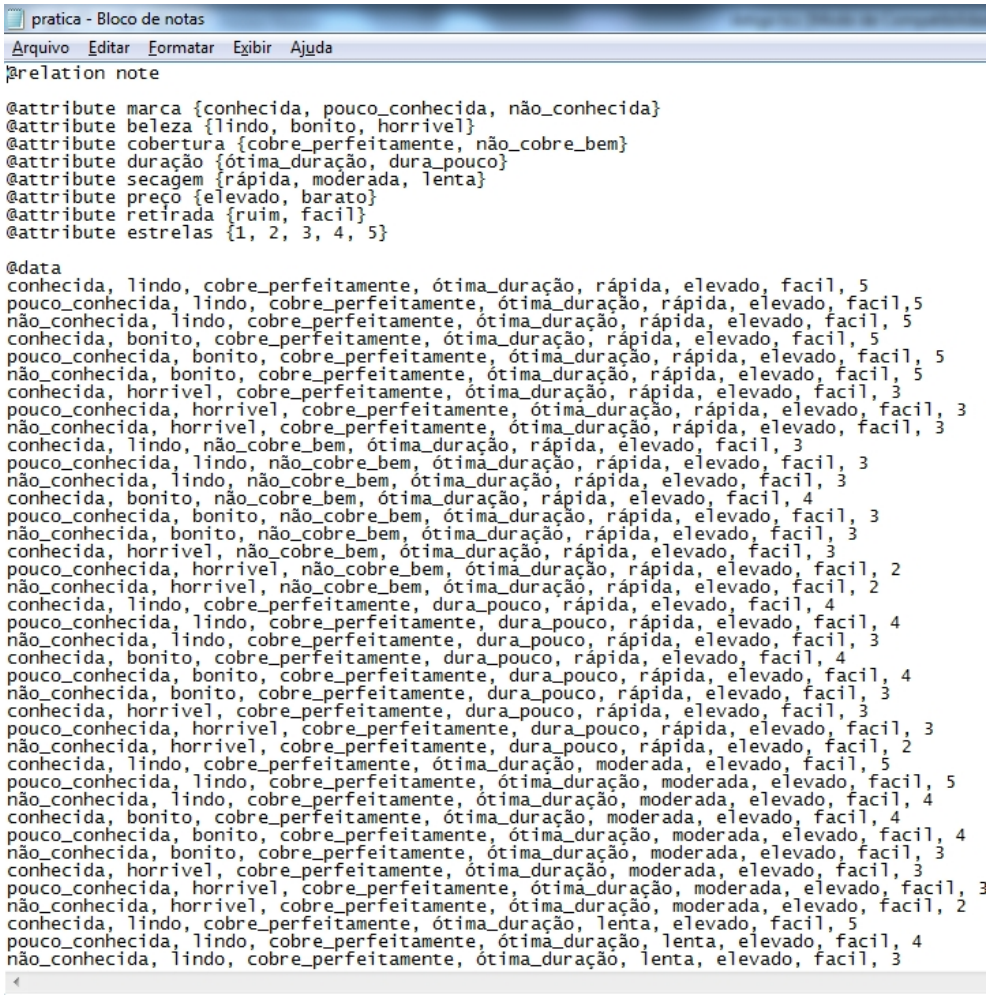


Figura 2. Arquivo no formato ARFF. **Fonte:** Dados do autor, 2015.

Tendo as opiniões organizadas, o passo seguinte é extrair das resenhas lidas elementos textuais, ou seja, palavras–chave que podem ser palavras ou pequenas expressões consideradas importantes para caracterizar o produto em estudo. Palavras-chave são termos que expressam sentimentos positivos ou negativos capazes de qualificar em produto ou objeto em análise [11].

Estes termos são definidos pelas avaliadoras do produto, no caso as blogueiras, como são conhecidas.

Após a seleção das palavras-chave o próximo passo é preparar os dados para o processamento, ou seja, gerar um arquivo na extensão do software utilizado para mineração de opiniões. Neste caso optou-se pelo Weka.

O Weka é um software que possui diversas implementações de algoritmos de aprendizagem de máquinas de diferentes paradigmas, além de algoritmos para preparação de dados e validação dos resultados. O ambiente foi desenvolvido na Universidade de Waikato na nova Zelândia, sendo todo escrito em Java e de código aberto disponível para uso, consulta e adequações por parte de seus usuários [12].

O Weka possui um formato de arquivo de entrada próprio "ARFF" (*atribute relato file forma*), basicamente dividido em duas partes: a primeira é composta por um cabeçalho que contém uma lista de todos os atributos e seus tipos e a segunda parte consiste das instâncias que serão utilizadas. Para transformar os dados extraídos das resenhas para este tipo de arquivo os passos a seguir devem ser executados [2]:

1. Abrir um arquivo como texto simples e salvar com extensão ARFF;
2. Rotular o conteúdo do arquivo para que a ferramenta saiba o que significa cada campo:

@relation – Usado para rotular o conjunto de dados

@attribute – usado para rotular atributos

@data usado para identificar os dados.

A ferramenta Weka é formada por um conjunto de pacotes: *atribute selection, classifiers, clustering, association rules e estimators.* Cada pacote é formado por vários algoritmos que possuem funções específicas de acordo com as tarefas de mineração de dados citadas no capitulo 2.2 deste trabalho [9].

Segundo (GUEDES, 2010 [11] e (SANTOS, 2008)[9] que fizeram uma aplicação semelhante de mineração, neste trabalho foi utilizado o algoritmo de classificação J48, pois para este tipo de estudo e o algoritmo mais indicado.

Tendo com dificuldade de representantes e empresários do ramo de cosméticos, selecionar produtos que possam atrair mais consumidores para seus negócios, foram retirados atributos que apareceram com mais frequência nas resenhas lidas em blogs de beleza e cosméticos sobre esmaltes, produto escolhido para exemplificar o uso da mineração por ser um dos produtos de cosméticos mais vendidos, e as palavras que mais apareceram foram as seguintes em ordem de mais importantes:

Marca: Segundo observado nas resenhas e dito em entrevista feita com proprietárias de blogs e representante de produtos de beleza, é o primeiro item a ser observado

pelo consumidor, dividida em conhecida, pouco conhecida, e não conhecida.

Beleza: referente à cor do esmalte dividida em lindo, bonito e horrível.

Cobertura: referente à transparência da cor do esmalte divida em cobre perfeitamente e não cobre bem.

Duração: indica o tempo que o produto dura na unha das clientes como ótima duração e dura pouco.

Secagem: referente ao tempo que o produto demora a secar este atributo foi divido em rápida, moderada e lenta.

Preço: referente ao custo do produto, elevado, barato.

Retirada: este atributo indica qual a facilidade de se retirar o produto da unha, dividido em ruim e fácil.

Estrelas: este atributo e referente à nota dada ao produto pelas avaliadoras, sendo em uma escala de 1, como ruim, a 5, como ótimo.

O próximo passo foi criar um arquivo na extensão ARFF, formato de arquivo reconhecido pelo software Weka.

A partir deste formato o arquivo ficou pronto para ser importado pelo software Weka.

APLICAÇÃO DO ALGORITIMO J48

O J48 constrói uma árvore de decisão. A forma de construção é uma abordagem *Top-down,* em que o atributo mais significativo, ou seja, o mais generalizado, quando comparado a outros atributos do conjunto, é considerado raiz da árvore. Na sequência da construção, o próximo nó da árvore será o segundo atributo mais significativo, e assim, sucessivamente até gerar o nó folha, que representa o atributo alvo da instancia [13].

Neste trabalho as principais métricas analisadas foram *precision, recall e f-measure e* o pacote utilizado será *classifiers,* pois são as principais métricas para a valiar os resultados neste tipo de aplicação [14].

Para um melhor entendimento, é necessário explicar alguns conceitos. Suponha a existencia de duas classes, a classe + (positivo) e a classe – (negativo). A forma de classificar objetos dentre essas duas classes são quatro :

• Verdadeiro positivo (VP): Número de objetos que foram classificados como sendo da classe + e que realmente pertencem à classe +;

• Falso negativo (FN): Número de objetos que foram classificados como sendo da classe – e que pertencem à classe +;

• Falso positivo (FP): Número de objetos que foram classificados como sendo da classe + e que pertencem a classe - ;

• Verdadeiro negativo (VN): Número de objetos que foram classificados como sendo da classe - e que realmente pertencem à classe -.

A métrica precision representa a relação de objetos positivos que foram identificados como positivo:

$$P = \frac{VP}{VP + FP}$$

A métrica recall representa a quantiade de objetos positivos que foram identificados corretamente:

$$R = \frac{VP}{VP + FN}$$

As métricas precision e recall podem ser resumida pela medida F-Measure, que pode ser calculada da seguinte forma:

$$F1 = \frac{2RP}{R + P} = \frac{2VP}{2VP + FP + FN}$$

3. RESULTADOS E DISCUSSÃO

Gerando um arquivo na extenção ARFF através do weka, contendo somente os registros equivalentes ao *classfy*, obtiva-se os seguintes resultados:

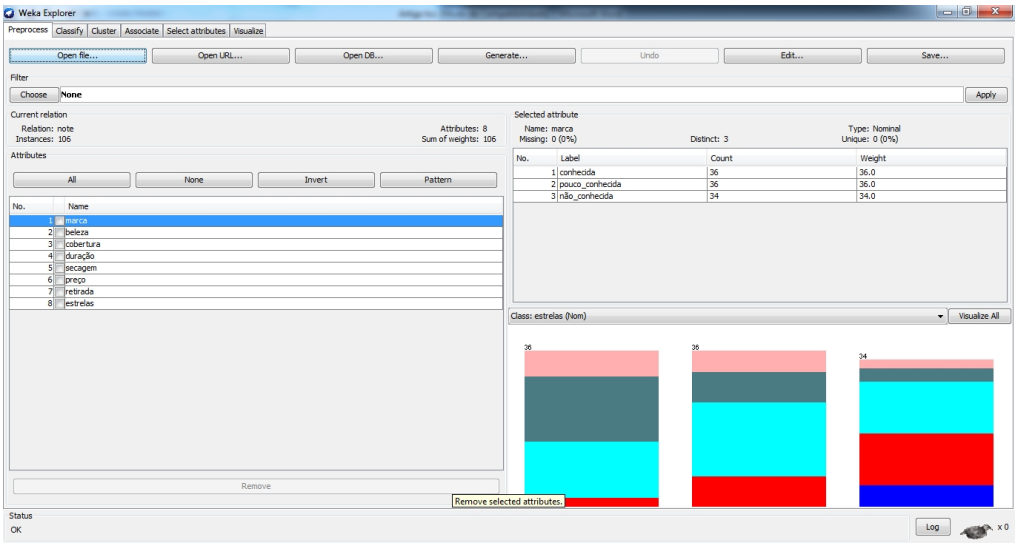


Figura 3. Arquivo ARFF obtido pelo software Weka no campo preprocess. **Fonte:** Dados do autor.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.8	0.04	0.5	0.8	0.615	0.61	0.974	0.491	1
	0.619	0.047	0.765	0.619	0.684	0.621	0.947	0.764	2
	0.738	0.172	0.738	0.738	0.738	0.566	0.877	0.788	3
	0.68	0.099	0.68	0.68	0.68	0.581	0.914	0.682	4
	0.923	0.022	0.857	0.923	0.889	0.873	0.991	0.907	5
Weighted Avg.	0.726	0.105	0.733	0.726	0.726	0.62	0.918	0.759	

Figura 4. Relatório gerado pelo processamento do algoritmo J48. **Fonte:** Dados do autor, 2015.

Observe, na classificação que, por exemplo, para a classe 5 estrelas a precisão foi de 85,70%, com *recall* de 92,30% e *F-Measure* de 88,90%.

A matriz de confusão contém informações importantes para o entendimento do resultado dado pelo algoritmo J48, como a quantidade de instâncias classificadas corretamente, a quantidade de instâncias classificadas erroneamente e a quantidade de instancias que o algoritmo acredita ser de um tipo, como por exemplo, 3 estrelas e na verdade foram classificados como 4 estrelas.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  4  1  0  0  0 |  a = 1  Somatória = 5
  2 13  5  1  0 |  b = 2  Somatória = 21
  2  3 31  6  0 |  c = 3  Somatória = 42
  0  0  6 17  2 |  d = 4  Somatória = 25
  0  0  0  1 12 |  e = 5  Somatória = 13
```

Figura 5. **Matriz de confusão.** Fonte: **Dados do autor**

É possível observar na primeira linha, por exemplo, que 4 instâncias foram classificadas corretamente como 1 estrela e uma instância foi classificada de maneira incorreta como 2 estrelas, e assim sucessivamente nas outras linhas da matriz de confusão.

Para saber como chegar ao resultado de instâncias classificadas corretamente e erroneamente basta analisar a imagem abaixo, no qual a somatória dos valores marcados em azul fazem parte dos valores classificados corretamente e os outros valores marcados em vermelhos nas outras diagonais, representam que foram classificados de forma incorreta.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  4  1  0  0  0 |  a = 1
  2 13  5  1  0 |  b = 2
  2  3 31  6  0 |  c = 3
  0  0  6 17  2 |  d = 4
  0  0  0  1 12 |  e = 5
```

Figura 6. Valores classificados corretamente e erroneamente na matriz de confusão. **Fonte:** Dados do autor, 2015.

Observando a figura acima obteve-se os seguintes resultados:

Instâncias classificadas corretamente: 77

Instâncias classificadas erroneamente: 29.

```
=== Summary ===

Correctly Classified Instances          77               72.6415 %
Incorrectly Classified Instances        29               27.3585 %
Kappa statistic                          0.6282
Mean absolute error                      0.1421
Root mean squared error                  0.2666
Relative absolute error                 48.4136 %
Root relative squared error             69.7213 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      37.3585 %
Total Number of Instances              106
```

Figura 7. Resultados. **Fonte:** Dados do autor, 2015.

A figura acima apresenta os casos classificados corretamente e incorretamente assim como as suas respectivas porcentagens. *Kappa statistic* é uma medida corrigida de acordo entre as classificações e as classes verdadeiras.

Observa se, por exemplo, que no atributo “retirada” quando igual a “ruim” a arvore gera o seguinte resultado: (3.0 / 1.0) que pode ser interpretado como: (3.0) que são as instâncias classificadas corretamente e (1.0) as instâncias classificadas de maneira incorreta.

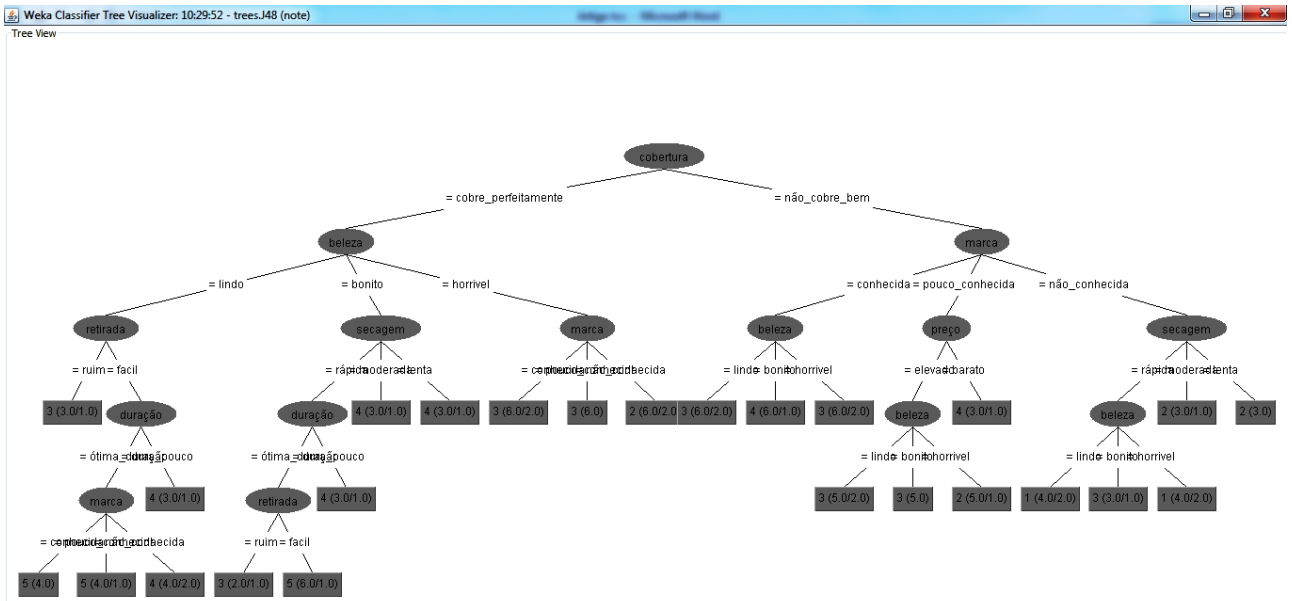
Na figura 8 pode-se observar a árvore de decisão criada pelo algoritmo:

4. CONCLUSÃO

Por meio da aplicação das técnicas de S*entiment analysis* e *Opinion mining*, foi possível a resolução do problema proposto para coleta e mineração de dados para seleção de produtos a serem vendidos ou representados. Os resultados foram satisfatórios, apesar das limitações e melhorias a serem feitas. Dentre elas destaca-se a necessidade da implementação de um programa para coletar as informações automaticamente de acordo com a necessidade do usuário, contribuindo para facilitar e agilizar pesquisas e o processo de analise de sentimentos.

Já na aplicação prática de mineração de dados os resultados também foram satisfatórios, já que o mesmo reduz custo de pesquisas de campo e traz informações de maneira que podem ser melhores visualizadas e interpretadas pelo usuário e assim analisar a melhor maneira de se tomar uma decisão na escolha de produtos a serem colocados nas prateleiras agrando e atraindo mais consumidores para o negócio.

Portando, a aplicação das técnicas de *Sentiment analysis* e *Opinion mining*, foi de grande aprendizado para o aluno que até então tinha tais técnicas apenas como



conceito, podendo utilizar tais técnicas na aplicação de conceitos de marketing futuramente.

.

Figura 8. Árvore de decisão. **Fonte:** Dados do autor, 2015.

REFERÊNCIAS

[01] Haaff M. Sentiment analysis hard but worth it, 2009. Disponível em: <http://www.customerthink.com/blog/sentiment_analysis_hard_but_worth_it>. Acesso em: 15 de junho de 2015.

[02] Frank E, Witten I. Data Mining Pratical Machine Learning Tools and Techniques. 2. ed. San Francisco: Elsevier. 2005.

[03] Ferreira, EBA. Análise de sentimento em redes sociais utilizando infuência das palavras, 2010. Disponível em: <http://www.cin.ufpe.br/~tg/2010-2/ebaf.pdf>. Acesso em 24 de Julho de 2015

[04] Oguri P. Aprendizado de Máquina para o Problema de Sentiment Classification. 2006. 54 f. Dissertação (Mestrado em informática) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro.

[05] Lima W. Sentiment Analisys: a nova corrida do ouro da informação digital, 2012. Disponivel em: <http://cio.com.br/opiniao/2012/04/05/sentiment-analysis-a-nova-corrida-do-ouro-da-informacao-digital/> Acesso em 09 de junho de 2015.

[06] Liu B. Web data mining: Exploring hiperlinks, Contents, and Use Data. 2. ed. Chicago: Springer, 2011.

[07] Amo S. Universidade Federal do Paraná. Disponível em: <http://inf.cp.utfpr.edu.br/ligia/papers/jai-cap5.pdf>. Acesso em 24 de Abril de 2015.

[08] Carvalho LA. Data mining - a mineraçãode dados no marketing, medicina, economia, engenharia e Administração. 1. ed. São Paulo: Érica. 2005.

[09] Santos DP. Mineração em notas fiscais de entrada de uma empresa calçadista. 2008. 83 f. Dissertação (Graduação em Ciência da Computação) – Instituto de ciências exatas e tecnológicas, Centro Universitário Feevale.

[10] Fuzzi LP. Oque é a pesquisa de campo, 2010. Disponivel em: <http://profludfuzzimetodologia.blogspot.com.br/2010/03/o-que-e-pesquisa-de-campo.html>. Acesso em: 01 de junho de 2015.

[11] Guedes R, Derkian A, Magalhães LH. Mineração de opiniões de usuário na busca de conhecimento. Faculdades integradas Vianna Junior. Juiz de Fora. 2010; 1(edição especial):84-99.

[12] Souza VMA, Feltrim VD. Análise automática de coerência semântica em resumos acadêmicos escritos em português, 2011. Disponivel em: <http://www.lbd.dcc.ufmg.br/colecoes/stil/2011/009.pdf > Acesso em 01 de Junho de 2015.

[13] Tavares C, Bozza D, Kono F. Descoberta do conhecimento aplicada a dados eleitorais. Gestão e conhecimento. São Carlos. 2007; 5(1):54-94.

[14] Jabour I. Análise estrutural para classificação de páginas na Web, 2008. Disponivel em: <http://www.puc-rio.br/pibic/relatorio_resumo2008/relatorios/ctc/inf/inf_iamj.pdf>. Acesso em 09 de junho de 2015.