8-10-2020

Opinion Mining in Social Media as a Tool to Measure Consumer (In)Satisfaction

Bel. Luis Sancliment Iglesias
Universidade Federal do Paraná – UFPR, luissiglesias@hotmail.com

Dr.ª Denise Fukumi Tsunoda
Universidade Federal do Paraná – UFPR, dtsunoda@ufpr.br

A Mineração de Opinião em Mídias Sociais como Ferramenta para Medir a (In)Satisfação do Consumidor

Artigo Completo

Bel. Luis Sancliment Iglesias
Bacharel em Gestão da Informação na Universidade Federal do Paraná – UFPR – luissiglesias@hotmail.com

Dr.ª Denise Fukumi Tsunoda
Professora da Universidade Federal do Paraná – UFPR - dtsunoda@ufpr.br

Abstract

Study that explores the contribution of opinion mining in customer service 2.0 databases extracted from Facebook, in order to measure consumers (in)satisfaction. The research aims to evaluate and propose tools used in the process of discovering knowledge in text and employing them in opinion mining at sentence level, as well as to use an opinion analysis methodology, choosing the NetVizz tool to extract the database from Facebook, the Microsoft Excel® for selection and reduction of data, Python codes for cleaning and transformation and the Semantria tool for text analysis. The pre-processed base for opinion mining is submitted to Naive Bayes, SMO and J48 algorithms in the Weka tool. The Research shows satisfactory results in opinion mining, with the best hit rate obtained in the SMO algorithm. Future works are proposed on customer service bases, for further comparison of obtained results on customer service 2.0 databases, seeking improvements.

Key-words

Sentiment Analysis. Text Mining. Opinion Mining. Customer Service 2.0. Customer Satisfaction. Consumer Satisfaction.

Resumo

Estudo que explora a contribuição da mineração de opinião em bases de dados SAC 2.0 extraídas do Facebook, para mensurar a (in)satisfação dos consumidores. A pesquisa visa avaliar e propor ferramentas utilizadas no processo de descoberta de conhecimento em texto e emprega-las na mineração de opinião a nível de sentença, bem como utilizar uma metodologia de análise de opinião, selecionando a ferramenta NetVizz para extrair a base de dados do Facebook, o Microsoft Excel® para seleção e redução de dados, códigos em Python para limpeza e transformação e a ferramenta Semantria para analisar o texto. Submete-se a base pré-processada para a mineração de opinião aos algoritmos *Naive Bayes*, SMO e J48 na ferramenta Weka. A pesquisa apresenta resultados satisfatórios na mineração de opinião com melhor taxa de acerto obtida no algoritmo SMO. Propõe-se trabalhos futuros em bases SAC, para posterior comparação dos resultados obtidos em bases SAC 2.0, buscando aprimoramentos.

Palavras-chave

Análise de Sentimentos. Mineração de Texto. Mineração de Opinião. SAC 2.0. Satisfação do Cliente. Satisfação do Consumidor.

Introdução

O advento das novas tecnologias de comunicação em proporções cada vez maiores, faz com que as mídias sociais sejam ferramentas que possibilitam a interação entre as pessoas, sejam elas conhecidas entre si ou não. Tais interações ocorrem quer seja para os indivíduos possam se pronunciar quanto a algum assunto que tenha sido do seu interesse ou para expressar suas opiniões e sentimentos a respeito dos mais diferentes assuntos. Estas opiniões podem expressar, por exemplo, a (in)satisfação quanto a produtos ou serviços de determinadas marcas que foram consumidos e também buscar obter informações, com o intuito de decidir sobre adquirir ou não determinados produtos e serviços, aliados a suas respectivas marcas.

As informações veiculadas nas mídias sociais podem ser valiosas para as organizações, uma vez que viabilizam maior aproximação com os consumidores e possibilitam não somente mensurar o grau de (in)satisfação dos mesmos, mas também responder de forma ágil às solicitações, questionamentos e reclamações efetuadas, bem como aumentar as chances de conquistar novos clientes.

Para este novo formato de interação cliente-empresa é que surgiu o SAC 2.0 (Serviço de Atendimento ao Consumidor 2.0) que segundo a agência wck (2016) é uma evolução do SAC tradicional que visa um atendimento mais completo, tendo o consumidor maior interação e voz ativa, de forma que o problema dele possa ser resolvido com maior rapidez e de maneira otimizada. A utilização da ferramenta SAC 2.0 é mais popular nas mídias sociais, e apesar das empresas ainda utilizarem o SAC tradicional, o consumidor moderno já está acostumado com uma forma mais rápida de interação.

Com o volume de dados em crescimento exponencial que circula nas mídias sociais, faz-se necessário que sejam utilizadas ferramentas que permitam a captura de texto em linguagem natural, para um posterior uso da mineração de opinião, que permitirá que sejam extraídos dados relevantes de forma automatizada, sumarizando em resultados que irão proporcionar uma melhor visualização e maior agilidade na tomada de decisão, resolução de problemas e mensuração da (in)satisfação dos consumidores.

Por conseguinte, a questão desta pesquisa é verificar: como a mineração de opinião em bases de dados extraídas de mídias sociais pode contribuir para a medição da (in)satisfação dos consumidores?

Os objetivos que darão o norte para poder responder à questão da pesquisa exposta são: estudar a mineração de opiniões e propor ferramentas que auxiliem na extração e pré-processamento de bases de dados retiradas de mídias sociais; escolher e aplicar ferramentas que possam ser utilizadas para a mineração de opiniões proposta e registrar a metodologia de análise de opiniões utilizada, sumarizando etapas, ferramentas e métodos.

Referencial teórico

A satisfação do consumidor é primordial para as organizações, por isso, para elas lograrem com que seus produtos e serviços sejam consumidos, é preciso primeiramente alcançar a satisfação de seus clientes. Os hábitos dos consumidores sofrem mudanças frequentes e com elas a forma com que eles pensam também. Segundo Kotler (2012), a satisfação tanto pode consistir em um sentimento de prazer como de decepção, tal sentimento resulta da comparação entre o desempenho de um produto e as expectativas do consumidor.

SAC 2.0

Com o surgimento da Internet e das mídias sociais, as empresas ante a necessidade de atender melhor o consumidor e como uma opção alternativa ao SAC tradicional por meio de telefone, se propuseram a procurar novas formas de dar suporte ao cliente, ouvir seus elogios, dúvidas, comentários, sugestões e reclamações. As organizações identificaram que as mídias sociais poderiam ser utilizadas para tal finalidade. Surgiu então o SAC 2.0, que nada mais é que um SAC voltado às mídias sociais. Segundo Gonsalves e First (2013), o SAC 2.0 constitui o serviço de atendimento ao consumidor nas mídias sociais, principalmente no Twitter e Facebook. Pelo fato da internet estar presente no dia a dia das pessoas, o SAC 2.0 vem sendo mais usado pelos consumidores, pois o ambiente proporciona rapidez e agilidade. O grande diferencial além da velocidade é que diferente do SAC tradicional que é praticamente de mão única, o SAC 2.0 é um canal de duas vias, onde existe uma interação entre os usuários e as organizações.

Indicadores de satisfação

Saber o quanto os clientes estão satisfeitos são medidas de grande utilidade para as organizações, elas auxiliam a descobrir o grau em que um produto ou serviço está cumprindo com as expectativas do cliente. Através do uso dos indicadores de satisfação, é possível determinar quais são os pontos favoráveis e onde precisam haver melhorias. Para Slack e Lewis (2002), existem basicamente cinco objetivos de desempenho que visam atender as exigências dos clientes e tem significado para qualquer tipo de operação, obedecendo prioridades diferentes, dependendo da situação: qualidade, rapidez, confiabilidade, flexibilidade e custo.

Mídias sociais

As mídias sociais são uma nova forma de comunicação por meio de estruturas que permitem entre outras coisas que pessoas e empresas possam interagir e trocar informações entre si, de forma rápida e em tempo real, seja por algum interesse mútuo ou até mesmo por querer demonstrar sua opinião. Telles (2010), alega que diversas pessoas se confundem com os termos "mídias sociais" e "redes sociais" e por várias vezes os usam de forma errônea. O autor afirma que os termos diferem, pois, redes sociais são uma categoria de mídias sociais. Segundo a Post Digital (2018), "uma rede social é focada na criação ou manutenção de relacionamentos entre as pessoas, e uma mídia social é mais focada no compartilhamento de conteúdo".

Segundo Franco (2018), as mídias sociais Youtube e Facebook estão entre as primeiras colocações no ranking dos sites mais usados no Brasil. Para Johnson (2014), o Facebook é uma espécie de serviço de rede social que foi lançado em 2004, e para que possa ser utilizado, faz-se necessário que o usuário se registre e crie seu perfil. Uma vez criados os perfis, os usuários podem fazer troca de mensagens, publicar seu status, publicar fotos e conversar com outros usuários via chat, entre outras funcionalidades. Com o tempo, a rede social converteu-se em uma rede mais profissional, com possibilidades de uso pelas organizações como a de informar e interagir com os consumidores.

Descoberta de conhecimento em texto (DCT)

As mídias sociais estão repletas de informações, dados, notícias, imagens, vídeos e em alguns casos mais específicos como o Facebook, há um conteúdo bem significativo em texto, tanto de pessoas como de organizações que escrevem mensagens contendo opiniões, comentários e reclamações dos mais variados assuntos. O grande desafio é que estes textos se encontram em linguagem natural, isto é, na linguagem razoavelmente inteligível para o ser humano, mas não para os computadores.

Para que possa haver uma interpretação desses textos pelas máquinas, é realizada a etapa de pré-processamento prévia à mineração denominada Descoberta de Conhecimento em Texto (DCT). Para Schiessl e Bräscher (2011), devido à complexidade da linguagem natural para a interpretação direta das máquinas, é necessário fazer uma extração de conhecimento das bases textuais e criar agrupamentos e modelos de classificação automatizados para que possam ser interpretados por computadores.

Mineração de opinião

Segundo Liu (2015), a mineração de opinião, é o estudo computacional das opiniões, sentimentos, atitudes e emoções das pessoas. A mineração de opinião é dirigida principalmente a opiniões que exprimem sentimentos positivos ou negativos. Deve-se considerar também expressões que não denotam nenhum sentimento, as denominadas expressões neutras. Além da opinião e do sentimento, existem os conceitos de afeto, emoção e humor, que vem a ser os estados psicológicos mentais. Conforme Liu (2015), a análise de sentimentos se conduz em três níveis. O primeiro nível denominado nível de documento (*document level*), onde se classifica primeiramente todo o documento para saber se está expressando um sentimento positivo ou negativo. O segundo nível é denominado nível de sentença (*sentence level*), onde é analisado se cada frase expressa uma opinião positiva, negativa ou neutra, que normalmente significa sem opinião. O terceiro nível denomina-se nível de aspecto (*aspect level*), onde diferentemente dos níveis um e dois, que em nenhum caso as análises denotam se as pessoas gostam precisamente ou não, no nível três a análise sim consegue diferenciar e examina diretamente a opinião e seu alvo.

Metodologia

Seguindo a classificação de Liu (2015), o nível de análise de sentimento que foi adotado nesta pesquisa é o segundo nível denominado nível de sentença, que visa a análise das frases de consumidores com opiniões de subjetividade positiva, negativa e neutra. Com este nível de análise, intenciona-se poder classificar as opiniões dos consumidores sobre produtos e serviços, considerando a satisfação em opiniões positivas e a insatisfação em opiniões negativas. Para as opiniões em que os consumidores não expressam claramente uma opinião, foram consideradas como opiniões neutras.

Bases de dados

A base de dados utilizada para esta pesquisa foi retirada da página oficial da empresa Ford Brasil no Facebook, especificamente dos comentários da postagem de 30 de julho de 2018, sobre a nova linha Ford Ka 2018 que contava quando foi capturada a base com 8 mil visualizações, 64 mil curtidas, 1.169 compartilhamentos e 1.114 comentários. O motivo de ter sido selecionada a base de dados sobre a postagem mencionada da empresa de automóveis Ford, é por contemplar nas postagens opiniões positivas, negativas e neutras de forma equilibrada, relevante para que os objetivos deste estudo sejam atingidos, uma vez que visam mensurar tanto as opiniões de satisfação como as de insatisfação. Outro motivo de ter sido selecionada uma empresa automobilística é por ela abranger não somente as opiniões do produto em si, mas também as de outros serviços e atendimentos agregados que estão envolvidos, como o de vendas, pós-vendas, revisões, garantia, entre outros.

Facebook e NetVizz

A extração da base de dados foi efetuada no Facebook, o fato de ter-se escolhido esta rede social para a pesquisa, deve-se a que a mesma consta em segundo lugar em mídias sociais no ranking de sites mais utilizados no Brasil, segundo a empresa Alexa (2018), este ranking é obtido pela média de visitantes que diariamente acessam um certo site e o número de visualizações deste mesmo site no período do último mês.

Na coleta de dados foi utilizada a aplicação NetVizz, que foi uma ferramenta criada por Bernhard Rieder em 2009 e permite extrair dados de seções da plataforma, gerando arquivos com extensão “tab” e possibilitando que sejam analisados para fins de pesquisa. Segundo Rieder (2013), a ferramenta inicialmente se desenvolveu com o intuito de estudar uma interface de programação de aplicações (API) para o Facebook como um novo objeto de mídia e para a avaliação de métodos nativos digitais.

Aspectos éticos da pesquisa

Por questões de segurança e ética quanto à privacidade de dados dos usuários, a ferramenta NetVizz possui limitações de extração de dados ocasionadas pelas restrições estabelecidas pelo Facebook. Assim mesmo, nesse estudo foi tomado o cuidado de excluir-se a identificação das pessoas para preservar sua privacidade.

Descoberta de conhecimento em texto

O método utilizado nesta pesquisa é baseado no processo de Descoberta de Conhecimento em Texto (DCT), que tem como objetivo a aquisição de conhecimento através das informações obtidas na transformação de forma automática de dados textuais. Após esta pré-seleção da base de dados, são realizadas as etapas de pré-processamento, mineração de opiniões e a avaliação (pós-processamento).

Segundo Lunardi, Viterbo e Bernardini (2015), a mineração de opiniões por meio de técnicas de aprendizado de máquina supervisionado, são as que têm maior utilidade para a definição de modelos para classificação de opiniões. Para esta pesquisa foram usados algoritmos que utilizam o aprendizado de máquina supervisionado, com o método de processamento de linguagem natural relacionado com análise de sentimentos dos consumidores quanto a produtos e serviços, nos conceitos positivo, negativo e neutro.

Pré-processamento

A fase do pré-processamento visa uma prévia preparação dos dados para serem submetidos ao processo de mineração. Segundo Neves (2003), o pré-processamento se divide em quatro etapas: entendimento dos

dados, seleção, limpeza e transformação. A etapa de entendimento abarca a análise dos dados coletados e a definição da sua importância e significado. A seleção abrange a definição dos dados que serão utilizados para serem submetidos à mineração e quais deverão ser descartados. A limpeza dos dados consiste em elevar a qualidade dos dados para níveis propícios para o processo de mineração. A etapa de transformação visa preparar os dados de forma a estarem adequados para a técnica de mineração que foi utilizada. As operações envolvidas nesta etapa são a normalização, conversão de valores, discretização e composição dos atributos (NEVES, 2003). A divisão do pré-processamento pode ser observada na Figura 1 abaixo.

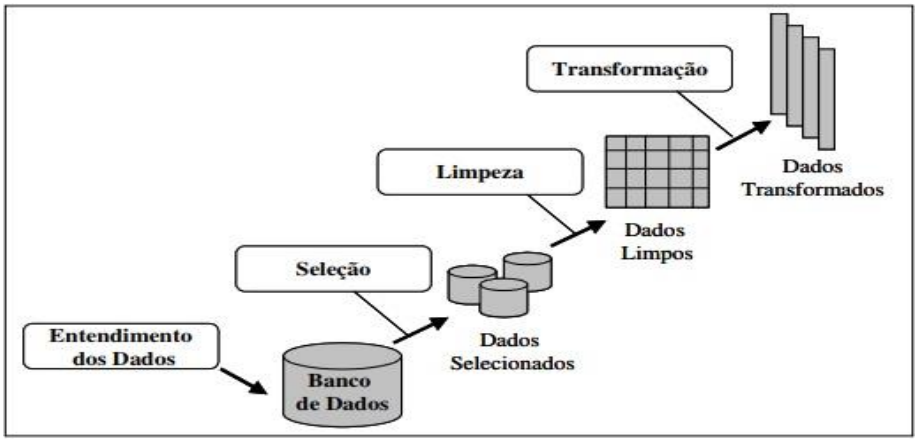

Figura 1. Divisão das etapas do pré-processamento de dados

Para a etapa de pré-processamento deste estudo, foram adotadas as seguintes ferramentas:

- **Microsoft Excel©:** foi utilizado inicialmente na etapa do pré-processamento para fazer a seleção dos dados e a redução da quantidade de objetos da base que será utilizada nesta pesquisa, bem como uma limpeza prévia de registros (linhas) em branco.
- **Python:** foi utilizado um código de programação em linguagem Python versão 3 por Rodrigues (2017). Este código tem a finalidade de efetuar: limpeza e transformação de dados, por padronização, que segundo Castro e Ferrari (2016), faz-se necessário para solucionar diferenças de unidades e escalas dos dados. Outra padronização necessária segundo os autores, é a de retirar caracteres especiais e de acentuação, que podem entrar em conflito com ferramentas de mineração em determinados idiomas.
- **Semantria:** para efetuar o processo de análise de texto foi utilizada a ferramenta Semantria na versão 2013 x64 6.0.102 que é uma solução de análise de texto e sentimento para pesquisas, mídias sociais e análises, desenvolvida pela empresa Lexalytics. Segundo a Lexalytics (2018), a análise de sentimentos é baseada no processo de determinar se partes textuais são positivas, negativas ou neutras. O site descreve que um sistema de análise de sentimentos para análise de texto combina o processamento de linguagem natural (PLN) e técnicas de aprendizado de máquina para atribuir pontuações de sentimento ponderadas às entidades, tópicos, temas e categorias dentro de uma sentença ou uma frase.

De forma a avaliar o desempenho da ferramenta Semantria (utilizada na última fase de pré-processamento), foram usados 3 outros códigos, que são variantes do código desenvolvido por Rodrigues (2017). A diferença está em que um dos códigos retira todas as *stopwords* (palavras irrelevantes) programadas exceto o "nao", outro código retira todas as *stopwords* programadas inclusive o "nao" e o terceiro código não retira as *stopwords*. Com o processamento destes três códigos em Python, foram geradas três bases de dados em arquivos "txt" mais uma quarta base de dados, também em arquivo "txt", sendo esta, com os dados brutos, sem nenhum tipo de tratamento do código em Python.

Processamento

A etapa de processamento é constituída pela mineração de opinião, que visa extrair conhecimento através do descobrimento de padrões e regras que possam atribuir significado nos dados que estão sendo submetidos. Para este estudo optou-se pela utilização do software Weka, com aplicação de algoritmos de classificação, por serem mais conhecidos e adequados na associação de dados com conjuntos de objetos que tem definição prévia de classes.

O Weka permite tarefas de pré-processamento, classificação, agrupamento e visualização dos dados através de sua interface gráfica. Castro e Ferrari (2016) afirmam ainda que é possível executar análises mais complexas através da criação de fluxogramas que concatenam tarefas de mineração de dados.

Como a base de dados utilizada para a mineração é composta por um atributo do tipo *PhraseSentiment* (numérico) e por um atributo do tipo *string* (sequência de caracteres), alguns algoritmos de mineração não podem ser utilizados diretamente, pois são incompatíveis com bases que contenham atributos com estas características. Por esse motivo os algoritmos considerados nesta pesquisa foram executados no software Weka com o uso de filtros, como é o caso do "*FilteredClassifier*" que habilita a discretização de atributos nominais vazios, atributos do tipo *string*, valores ausentes, atributos relacionais, atributos binários, numéricos e nominais, entre outros. Também foi utilizado o filtro não supervisionado "*StringToWordVector*" que converte os atributos *string* em um conjunto de atributos que representam informações de ocorrências de palavras. Adicionalmente também foi utilizado um "*tokenizer*" que é uma operação que permite dividir uma sequência de *strings* em partes como: frases, palavras, símbolos e outros elementos denominados "*tokens*" para aprimorar a mineração de opinião.

Os algoritmos escolhidos para a mineração da base selecionada dentro do software Weka foram o *Naive Bayes* que usa o modelo de classificação probabilístico, o SMO (*Sequential Mininal Optimization*) que utiliza o modelo baseado em função e o J48 que é um algoritmo baseado em árvores de decisão. Estes algoritmos foram selecionados devido a serem mais comumente utilizados e mais adequados ao tipo de base de dados que está sendo analisada.

No processo de configuração e execução dos algoritmos no Weka, foi utilizada a base em formato "arff" já previamente pré-processada no software Semantria.

Pós-processamento

A fase de pós-processamento visa explicar as formas de avaliação dos resultados obtidos na análise efetuada com o software Weka com os três algoritmos. Os resultados que foram utilizados neste estudo são:

- modelo de classificação (*classifier model*): apresenta os resultados e as representações textuais dos modelos de classificação que foram usados nos dados do treinamento em cada um dos algoritmos;
- estatística Kappa (*Kappa statistic*): medida estatística que tem por finalidade verificar o grau de confiabilidade intermediária. O seu valor indica quanto os dados coletados são representações corretas das variáveis. Segundo Landis e Koch (1977), uma possível interpretação dos valores seria: (a) deficiente: menor que 0,21; (b) justo: de 0,21 a 0,40; (c) moderado: de 0,41 a 0,60; (d) substancial: de 0,61 a 0,80 e (e) perfeito: de 0,81 a 1,00;
- taxa de acerto (*classified instances*): quantidade e percentual de instâncias corretamente e incorretamente classificadas pelo algoritmo;
- matriz de confusão (*confusion matrix*): matriz que demonstra o número de classificações reais comparadas com as classificações preditas de cada classe. Assim, pode ser verificado quantas foram classificadas de forma correta e quantas de forma incorreta para cada classe.

Resultados

Foram executadas no *software* Semantria as quatro bases de dados geradas após o pré-processamento no código Python conforme detalhado na seção de metodologia. O Semantria efetuou diversas análises, e dos resultados compatíveis com o estudo desta pesquisa, foi selecionado o "*EntityThemeDetail*" pois categoriza a frase analisada separando a "entidade", o "tipo de entidade" e o segmento da frase que caracteriza o sentimento, para depois classificá-lo como negativo, positivo ou neutro. Os campos "*HighlightedText*" e "*EntityThemeSentiment*" foram utilizados para gerar as bases com extensão "arff" e submetidas para análise no software Weka na etapa de processamento. Basicamente o campo "*HighlightedText*" contém o texto analisado, e o campo "*EntityThemeSentiment*" o sentimento (positivo, neutro ou negativo), percebido pelo Semantria. Outros resultados gerados pela ferramenta Semantria foram as nuvens de palavras, que apresentam a análise de frequência dos termos para cada uma das bases, conforme o tratamento a que foram submetidas.

Resultados do processamento das bases analisadas

Os experimentos em todas as bases analisadas foram realizados no software Weka com validação cruzada de 10 participações, com a utilização dos filtros "*FilteredClassifier*" e "*StringToWordVector*" e com o uso de "*tokenizer*". As taxas de acerto obtidas foram mais altas no algoritmo SMO para todas as 4 bases submetidas. Foi possível verificar que o número de instâncias (registros) geradas pelo *software* Semantria para cada base foi diferente. Isso foi devido à análise que o Semantria executa em cada base, com a influência do tratamento que é efetuado na etapa do pré-processamento com os códigos Python, onde é feita a limpeza e a transformação dos dados.

As informações de número de instâncias, atributos e os resultados obtidos para cada um dos sentimentos das bases analisadas quanto aos registros classificados corretamente para cada algoritmo, podem ser visualizados na Tabela 1.

Base de Dados	Nº Instâncias (registros)	Nº Atributos (campos)	Algoritmo *Naive Bayes (Nº instâncias classificadas corretamente)*			Algoritmo *SMO (instâncias classificadas corretamente)*			Algoritmo *J48 (instâncias classificadas corretamente)*		
			Positivas	Negativas	Neutras	Positivas	Negativas	Neutras	Positivas	Negativas	Neutras
Ford_Dados_Brutos	430	2	61	86	143	67	94	154	60	88	146
Ford_Com_StopWords	1241	2	114	168	664	139	201	819	113	161	802
Ford_Sem_StopWords_sem_nao	806	2	106	154	303	126	207	343	107	197	324
Ford_Sem_StopWords_com_nao	843	2	114	127	337	137	169	430	107	130	409

Tabela 1. Número de instâncias classificadas corretamente para cada algoritmo

Comparação dos resultados entre as bases

Com relação aos resultados obtidos entre as bases, é possível verificar que o algoritmo SMO obteve melhores resultados de taxas de acertos nas 4 bases submetidas para análise. Para a base bruta o percentual de taxa de acerto de instâncias classificadas corretamente foi de 73,3% utilizando o algoritmo SMO, 68,4% com o algoritmo J48 e 67,4% com o algoritmo *Naive Bayes*.

Para a base em que se mantiveram os *stopwords* o percentual de taxa de acerto para o algoritmo SMO foi de 93,4%, 86,7% com o algoritmo J48 e 76,2% com o algoritmo *Naive Bayes*. Para a base à qual foram retirados os *stopwords* inclusive o "nao" os percentuais de taxa de acerto foram de 83,9% utilizando o algoritmo SMO, 77,9% com o algoritmo J48 e 69,9% com o algoritmo *Naive Bayes*.

Para a base à qual foram retirados os *stopwords* exceto o "nao" os percentuais de taxas de acerto foram 87,3% utilizando o algoritmo SMO, 76,6% com o algoritmo J48 e 68,6% utilizando o algoritmo *Naive Bayes*.

Como pode ser visto na Tabela 2, o algoritmo SMO teve maior percentual de classificação correta em todas as situações, seguido do algoritmo J48. O algoritmo *Naive Bayes* foi o que apresentou menores percentuais de taxa de acerto em todos os experimentos realizados.

Base de Dados	Percentual de Classificações Corretas (%)		
	Algoritmo Naive Bayes	Algoritmo SMO	Algoritmo J48
Ford_Dados_Brutos	67,4%	73,3%	68,4%
Ford_Com_StopWords	76,2%	93,4%	86,7%
Ford_Sem_StopWords_sem_nao	69,9%	83,9%	77,9%
Ford_Sem_StopWords_com_nao	68,6%	87,3%	76,6%

Tabela 2. Percentual de classificação correta entre algoritmos

A respeito dos resultados de grau de confiabilidade intermediária atribuído pela estatística Kappa, para a base de dados bruta o grau encontrado foi "moderado" para os 3 algoritmos. Para a base com os *stopwords*

o grau foi "perfeito" para o algoritmo SMO, "substancial" para o algoritmo J48 e "moderado para o algoritmo *Naive Bayes*. Para a base que teve a retirada de *stopwords* inclusive o "nao" o grau encontrado foi "substancial" para os algoritmos SMO e J48 e "moderado" para o algoritmo *Naive Bayes*. Para a base que teve a retirada de *stopwords* excetuando o "não", o grau foi "substancial" unicamente para o algoritmo SMO e "moderado" para os algoritmos J48 e *Naive Bayes*.

Indicadores de satisfação

Utilizando os termos negativos e positivos classificados na análise do Semantria, é possível verificar a frequência das palavras e a presença de alguns objetivos de desempenho quanto a indicadores de (in)satisfação do consumidor. Também é possível estratificar palavras-chave, que podem denotar em ordem de frequência os motivos principais da (in)satisfação

Tomando como exemplo a Base Ford_Dados_Brutos que gerou 430 instâncias (registros), sendo 137 negativas, 196 neutras e 97 positivas, foi criada uma nuvem de dados para os 137 termos negativos. As maiores densidades foram encontradas nas palavras: "problema" com 27 ocorrências (19,7%), "problemas" com 7 ocorrências (5%), "defeito" com 6 ocorrências (4%), "reclamações" com 5 ocorrências (3%), "falta" com 4 ocorrências (3%), "péssimo" com 3 ocorrências (2%) e "prejuízo" com 3 ocorrências (2%). Destes 7 termos com maior frequência, 2 deles denotam relação direta com objetivos de desempenho: "defeito", atrelado ao objetivo qualidade, que se relaciona com a não conformidade de especificações de produto e "prejuízo", que pode estar relacionado com custo, que segundo Slack e Lewis (2002), também é um dos cinco objetivos de desempenho nos indicadores de satisfação.

Considerando que devem ser evitados os insatisfatores, mas também fortalecidos os satisfatores que motivam a compra, foi realizada a análise dos termos positivos. Assim, com respeito aos 97 termos positivos classificados pelo Semantria, utilizando também a base Ford_Dados_Brutos, foi gerada uma nuvem de palavras, das quais os 7 termos com maior densidade foram: "lindo" com 16 ocorrências (6%), "parabéns" com 11 ocorrências (4%), "melhor" com 10 ocorrências (4%), "sonho" com 10 ocorrências (4%), ótimo com 8 ocorrências (3%), "gostei" com 7 ocorrências (3%) e "confortável" com 7 ocorrências (3%).

Fluxo da metodologia utilizada na análise de opiniões

Para ter uma visão do processo compreendido neste estudo foi elaborado um fluxo com as etapas envolvidas na mineração de opinião e as ferramentas utilizadas em cada etapa com as principais funções dentro do processo. Este fluxo pode ser visualizado na Figura 2.

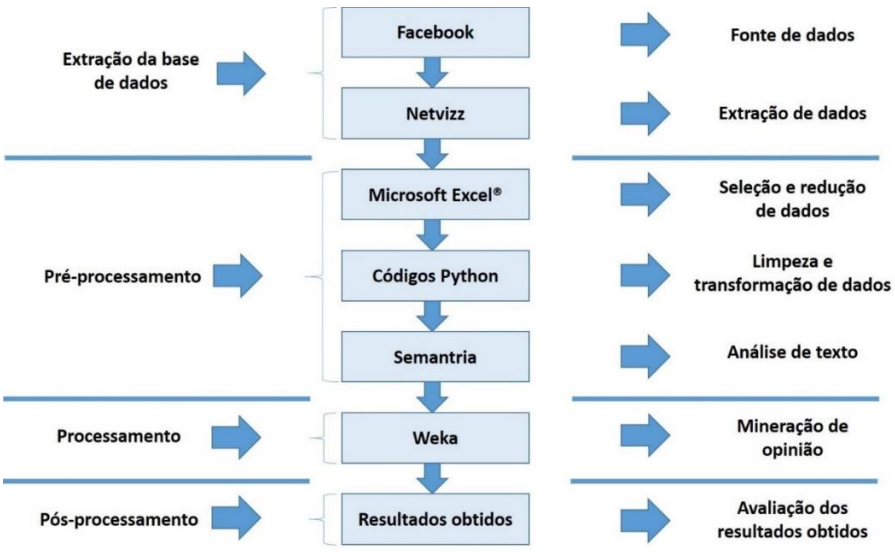

Figura 2. Fluxo da metodologia utilizada na análise de opiniões

Considerações finais

A pesquisa utiliza uma base do SAC 2.0, que contém opiniões, reclamações e sugestões de usuários nas mídias sociais, mais especificamente no Facebook, que tem uma ampla utilização por parte dos usuários. A base de dados foi extraída da página oficial da empresa Ford Brasil no Facebook, nos comentários efetuados na postagem de lançamento da nova linha Ford Ka 2018. A seleção desta base foi devido a que a mesma contém nas postagens opiniões positivas, negativas e neutras de maneira equilibrada, o que foi ideal para os objetivos deste estudo de avaliar tanto opiniões de satisfação como de insatisfação. O fato de ter sido selecionada uma empresa do ramo automobilístico, foi por ela compreender não somente as opiniões do produto em si, mas também de outros serviços e atendimentos envolvidos, como vendas, pós-vendas, revisões e garantia.

Um dos grandes desafios foi definir uma forma de tratamento de texto para transformar a linguagem natural utilizada nas mídias sociais em agrupamentos de textos, que pudessem ser interpretados e submetidos à mineração de opinião.

Objetivando-se encontrar ferramentas que pudessem auxiliar na extração e pré-processamento de bases de dados retiradas de mídias sociais, foram selecionadas as ferramentas Microsoft Excel®, códigos em Python e o Semantria, que em conjunto complementaram-se, cumprindo cada uma as diferentes tarefas de pré-processamento, tais como: selecionar e reduzir dados, limpeza e tratamento dos mesmos e finalizando com a análise de texto, para poder classificar automaticamente os comentários extraídos do Facebook em positivos, neutros e negativos.

Para efetuar a mineração de opiniões, foi selecionada a ferramenta Weka já utilizada em estudos e pesquisas que envolvem a mineração de opinião, e atendeu à finalidade por meio dos algoritmos de classificação *Naive Bayes*, SMO e o J48.

A partir da definição do local onde seria extraída a base e filtrando-se o segmento de dados da postagem que seria estudado, foram selecionadas as ferramentas em cada uma das etapas de extração de dados, pré-processamento, processamento e pós-processamento. Cada etapa foi realizada com pouca intervenção humana, e à medida que se conseguia sucesso nas etapas, foi possível verificar que cada ferramenta selecionada atingiu o esperado e preparou a base para a próxima ferramenta atuar. Ao término do processo, foi possível apresentar um fluxo com a sequência e respectiva ferramenta para a análise de opiniões.

Uma vez que a quantidade de dados nas redes sociais tem crescimento constante, a dificuldade de pessoas poderem explorá-las é cada vez maior, a utilização de máquinas e processamento de linguagem natural por meio do uso da mineração de opinião pode ser uma alternativa útil para mensurar a (in)satisfação dos consumidores de forma mais ágil e prática. Este estudo constatou que com a adoção de técnicas de pré-processamento é possível preparar bases de dados de forma adequada para a mineração de opinião, que irá criar agrupamentos de opiniões positivas, neutras e negativas e gerar a descoberta de padrões por meio de treinamentos efetuados por algoritmos de mineração adequados, e com isso, auxiliar no processo de gestão e atribuir maior agilidade na tomada de decisão. Com os resultados obtidos nesta pesquisa é possível afirmar que a mineração de opinião em bases de dados extraídas de mídias sociais pode contribuir na mensuração da (in)satisfação dos consumidores.

Para trabalhos futuros, recomenda-se a aplicação deste estudo utilizando-se uma base de dados SAC, onde existem algumas características distintas das encontradas em bases SAC 2.0 como linguagens mais padronizadas (linguagens documentais) e registros textuais que podem permitir um menor esforço de pré-processamento por serem efetuados por profissionais especializados e treinados para conduzir esta tarefa. Desta forma, poderiam haver comparações entre ambos os resultados dos estudos das bases e assim retirarem-se conclusões mais específicas e com elas a criação de metodologias mais aprimoradas.

REFERÊNCIAS

Agência WCK. 2016. “O que é SAC 2.0 e por que a sua empresa deve ficar atenta a ele?,” Disponível em: <https://agenciawck.com.br/o-que-e-sac-2-0-e-por-que-a-sua-empresa-deve-ficar-atenta-a-ele/>. Acesso em: 05 mai. 2020.

Alexa. 2018. “Alexa top sites: detailed description,” Disponível em: <https://www.alexa.com/topsites/countries/BR>. Acesso em 13 mai. 2020.

Castro, L. N.; Ferrari, D. G. 2016. "Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações," São Paulo: Saraiva.

Franco, A. H. C. 2018. "Inteligência coletiva: manifestações nos ambientes digitais," 141 f. Tese (Doutorado em CI) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista. Disponível em: <https://repositorio.unesp.br/bitstream/handle/11449/152741/franco_ahc_dr_mar.pdf?sequence=3>. Acesso em: 13 out. 2019.

Gonsalves, A; First, L. 2013. "SAC 2.0 como parte do planejamento estratégico de comunicação," 56 f. Trabalho de Graduação (Disciplina Tema Final) - Curso de Comunicação Social, Setor de Artes, Comunicação Social e Design, Universidade Federal do Paraná, Curitiba, 2013. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/52858/TCC_sac_2.0_como_parte_do_planejamento_estrategico_de_comunicacao.pdf?sequence=1>. Acesso em: 15 mai. 2020.

Johnson, A. J. 2014. "Su éxito en redes sociales," E-book. Disponível em: <https://pt.scribd.com/document/313463715/Su-Exito-en-Redes-Sociales-Amanda-J-Johnson>. Acesso em 13 mai. 2020.

Kotler, P. 2002. "Administração de marketing," 10. ed. São Paulo: Afiliada.

Landis J. R., Koch G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," 1 (33). Biometrics: 159–74.

Lexalytics. 2018. "Semantria for Excel," Disponível em: <https://www.lexalytics.com/semantria/excel>. Acesso em: 08 out. 2019.

Liu, B. 2015. "Sentiment Analysis: Mining Opinions, Sentiments and Emotions," New York: Cambridge University.

Lunardi, A. C.; Viterbo, J.; Bernardini, F. C. 2015. "Um levantamento do uso de algoritmos de aprendizado supervisionado em mineração de opiniões," Proceedings of XII Encontro Nacional De Inteligência Artificial E Computacional, 12., Natal. Anais... Natal: ENIAC, pp. 262-269. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2015/039.pdf>. Acesso em 22 out. 2019.

Neves, R. C. D. 2003. "Pré-processamento no processo de descoberta de conhecimento em banco de dados," 137 f. Dissertação (Mestrado) - Programa de Pós-graduação em Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul. Disponível em: < http://www.lume.ufrgs.br/bitstream/handle/10183/2701/000375412.pdf?sequence=1>. Acesso em: 28 mai. 2020.

Post Digital. 2018. "Qual a diferença entre rede social e mídia social?". Disponível em: <http://www.postdigital.cc/blog/artigo/qual-a-diferenca-entre-rede-social-e-midia-social#04>. Acesso em: 24 out. 2019.

Rieder, B. 2013. "Studying Facebook via data extraction: the Netvizz application," in *Annual Acm Web Science Conference*, 5., New York. Proceedings... New York: ACM, pp. 346-355. Disponível em: <http://thepoliticsofsystems.net/permafiles/rieder_websci.pdf>. Acesso em: 27 mai. 2020.

Rodrigues, A. C. F. 2017. "Modelo para análise de sentimentos no Facebook: um estudo de caso na página do senado federal brasileiro," 83 f. TCC (Graduação) – Curso de Gestão da Informação, Universidade Federal do Paraná, Curitiba. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/54864/Alan%20Cristian%20Falcoski%20Rodrigues.pdf?sequence=1&isAllowed=y>. Acesso em: 11 out. 2019.

Schiessl, M.; Bräscher, M. 2011. "Descoberta de conhecimento em texto aplicada a um sistema de atendimento ao consumidor," Revista Ibero-Americana de Ciência da Informação, v. 4, n. 2. Disponível em: <http://periodicos.unb.br/ojs311/index.php/RICI/article/view/1682/1481>. Acesso em: 22 out. 2019.

Slack, N; Lewis, M. 2002. "Operations Strategy," Harlow: Pearson Education.

Telles, A. 2010. "A revolução das mídias sociais: cases, conceitos, dicas e ferramentas," São Paulo, M.Books do Brasil Editora Ltda.