UNIVERSIDADE FEEVALE

FERNANDO AUGUSTO SCHUCH

DETECÇÃO AUTOMÁTICA DE SPAMS DE OPINIÃO EM AVALIAÇÕES DE PRODUTOS NA LÍNGUA PORTUGUESA

Novo Hamburgo

2019

FERNANDO AUGUSTO SCHUCH

DETECÇÃO AUTOMÁTICA DE SPAMS DE OPINIÃO EM AVALIAÇÕES DE PRODUTOS NA LÍNGUA PORTUGUESA

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pela Universidade Feevale

Orientador: Prof. Dr. Rodrigo Rafael Villarreal Goulart

Novo Hamburgo

2019

AGRADECIMENTOS

Agradeço a todas as pessoas que contribuíram para a realização desse trabalho de conclusão, em especial:

Aos meus pais, Dorian e Gaspar, por tudo o que fizeram por mim. Pela educação, amor, carinho e oportunidades de estudar e conquistar os meus sonhos.

À minha parceira, Catiele, por ser compreensiva e me dar forças para continuar no caminho certo, mesmo nos dias mais difíceis.

Ao meu orientador, Rodrigo, pelo suporte dado e sugestões que me ajudaram no desenvolvimento deste trabalho.

Aos voluntários que participaram da pesquisa, por me ajudarem a completar mais essa etapa. Somente através dessas contribuições, meu objetivo pôde ser alcançado.

Aos demais, que de alguma forma, estiveram ao meu lado até este momento.

Obrigado.

RESUMO

Opiniões sobre bens ou serviços representam uma excelente fonte de informação, tanto para consumidores quanto empresas e fabricantes. Avaliações sobre produtos em *websites* de venda estão sendo cada vez mais consultadas, com o propósito de tomar decisões de compra com base em experiências de outras pessoas. A confiança nessas avaliações é alta, principalmente em indivíduos entre 18 e 34 anos. Logo, percebe-se que há interesse e necessidade em estudá-las, a fim de acompanhar como está a reputação da marca na Internet. Pelo fato de que *reviews* positivas geralmente significam lucro, enquanto negativas afetam a notoriedade dos produtos, este cenário motiva a postagem de opiniões falsas, buscando persuadir consumidores a tomarem decisões erradas. Essa atividade, conhecida como spams de opinião, é uma vertente da mineração de opiniões que recebeu atenção somente a partir de 2008. Apesar de já existirem estudos nessa área, ela tem sido pouco abordada na língua portuguesa. Portanto, há escassez de exemplos anotados, ou seja, classificados como spam ou não-spam para a criação de algoritmos de detecção. Esta pesquisa apresenta o processo de anotação de um corpus de avaliações sobre mercadorias, objetivando a identificação de opiniões nas quais os usuários não têm experiência prévia com os produtos, analisando assim, o impacto delas na reputação online de bens de consumo. Verificou-se que em 29% das opiniões não foi possível afirmar se o indivíduo que a postou possuía a mercadoria, ou pelo menos a utilizou. Portanto, deixando dúvidas se sua avaliação é genuína. Com a utilização de algoritmos de Aprendizado de Máquina, foi proposto um classificador a partir de atributos linguísticos extraídos das opiniões. Com o experimento, concluiu-se que o modelo obteve 81% de acerto na predição de textos legítimos e possíveis spams.

Palavras-chave: Spams de opinião. Anotação. Processamento de Linguagem Natural. Aprendizado de Máquina.

ABSTRACT

Opinions about goods and services represent and excelent source of information, for both consumers and manufacturers. Evaluations about products on sales websites are increasingly being consulted, with the purpose of making purchase decisions relying on the experience of other people. Currently, with the use of the Internet on the rise, this practice has been adopted by many users. The confidence in these evaluations is high, mainly among individuals between 18 and 34 years old. Therefore, one can realize that there is interest and need to study them, in order to follow how well is the brand's reputation doing on the Internet. By the fact that positive reviews usually mean profit, negatives afect the products' reputation. This cenario motivates fake opinions posting, aiming to persuade consumers to make wrong decisions. This activity, known as opinion spam, is a branch of opinion mining, which received atention only after 2008. Despite the studies in this area, few have been done for Portuguese language. So, there is shortage of anotated examples, that is, classified as spam or non spam for the development of detection algorithms. This research presents the annotation process on a corpus of merchandises' reviews, with the goal os identifying opinions in which users do not have previous experience with the products, analysing their impact on the online reputation of consumer goods. It has been verified that in 29% of opinions it was not possible to guarantee the individual who posted it had the product, or at least used it. Therefore, leaving doubts if the evaluation was genuine. Using Machine Learning algorithms, it has been proposed a classifier with basis on linguistic attributes extracted from opinions. With this experimento, it has been concluded that the model got 81% of accuracy in predicting legitimate texts.

Keywords: Opinion spams. Annotation. Natural Language Processing. Machine Learning.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AMT	Amazon *Mechanical Turk*
AVP	Amazon *Verified Purchase*
IR	*Information Retrieval*
ML	*Machine Learning*
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
RBF	*Radio Basis Function*
RF	*Random Forest*
RSBL	Rede Social Baseada em Localização
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency – Inverse Document Frequency*
WEKA	*Waikato Environment for Knowledge Analysis*
WWW	*World Wide* Web
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Na última década, devido à evolução da Internet e crescimento no uso de *smartphones*, as redes sociais impulsionaram a geração de opiniões sobre produtos, serviços, estabelecimentos, pessoas e fatos. Atualmente, há uma grande facilidade em publicar e consultar tais informações nesse meio (CARDOSO; ALMEIDA, 2017). Segundo Jindal e Liu (2008), há um crescente interesse em minerar opiniões na Web, já que esse tipo de conteúdo contém informações valiosas para diversas aplicações.

As opiniões em mídias sociais têm sido utilizadas para diversos fins, desde decisões de compra, marketing e design de produtos, até escolhas em eleições políticas (LIU, 2012). De acordo com Lin et al. (2014), consumidores passaram a optar por determinados produtos baseando-se em experiências postadas por outras pessoas em *websites.* Esse comportamento motiva comerciantes a criar avaliações falsas, tanto para melhorar sua reputação quanto para atacar competidores.

Mukherjee, Liu e Glance (2012) complementam que se alguém quiser comprar um produto, irá ler suas *reviews*. Se a maioria for positiva, tenderá a comprá-lo, mas se grande parte for negativa, provavelmente o consumidor escolherá outra mercadoria. Esse cenário “proporciona fortes incentivos para a atividade conhecida como spams de opinião” (MUKHERJEE; LIU; GLANCE, 2012, p. 1, tradução nossa).

Recentemente, os números sobre a confiança em *reviews* online disponibilizados por Murphy (2018), afirmam que 91% das pessoas entre 18 e 34 anos, entrevistadas no Reino Unido, confiam nessas opiniões tanto quanto recomendações pessoais. Com isso, comerciantes sentem-se estimulados a postarem opiniões positivas sobre suas mercadorias, bem como desqualificar as de concorrentes (LIN et al., 2014). Devido à essa prática, organizações já se preocupam em criar esforços para reduzir o número de spams e *spammers*, ou seja, usuários que escrevem opiniões falsas. Em 2015, por exemplo, a Amazon entrou com ações judiciais contra sites que vendiam *reviews* falsas para comerciantes que anunciavam seus produtos no comércio eletrônico da empresa (BISHOP, 2015).

Spams de opinião, de acordo com Liu (2012), refere-se a atividade humana de enganar leitores ao escrever *reviews* injustas sobre determinados alvos. Embora fraudulentas, tais avaliações são cuidadosamente criadas com o intuito de parecerem

verdadeiras, o que dificulta a identificação pela simples leitura (LIU, 2012). Logo, spams inseridos dentro de opiniões postadas nas redes sociais podem afetar a experiência de consumidores, visto que os mesmos podem tomar decisões de compra ruins e consequentemente, perder a confiança em *reviews* online (SANDULESCU; ESTER, 2015).

Levando em consideração que pessoas têm dificuldade em detectar opiniões falsas através da leitura (LIU, 2012), além do grande número de *reviews* postadas diariamente (CONDORI; PARDO, 2017), diversas pesquisas foram propostas com o intuito de identificar spams de opinião de maneira automática. A maior parte delas, contudo, a partir de exemplos previamente obtidos na língua inglesa. Entretanto, a língua portuguesa carece de estudos na área. Até o momento, apenas Costa, Benevenuto e Merschmann (2013) investigaram o problema neste idioma.

Tendo isso em vista, a presente pesquisa visa complementar os trabalhos existentes, classificando *reviews* em termos de experiência do usuário com o produto. Significa identificar se o postador tem ou não experiência com a mercadoria avaliada por ter expressado que possui ou não o produto, ter utilizado ou, até mesmo conhece alguém que o tenha feito. Apesar desses tipos de opinião não necessariamente representarem que os indivíduos deliberadamente tentaram favorecer ou prejudicar seus alvos, Liu (2012) argumenta que, por não possuírem experiência prévia com o produto, sua opinião não é genuína.

Esta categoria de opinião, embora represente apenas uma fração do problema identificado por Jindal e Liu (2008), ainda é um obstáculo que pode estar afetando consumidores na hora de adquirir produtos. Portanto, é visível a necessidade de aumentar os esforços com o intuito de diminuir o número de spams presentes em *websites*.

Para atender a problemática da pesquisa, este trabalho apresenta uma metodologia para anotação manual de textos em português. Ela foi empregada por voluntários para classificação de conjuntos de textos reais, retirados de um site brasileiro de compra e venda de mercadorias. Além disso, a confiança na etiquetação é aferida por métodos aplicados para calcular o nível de concordância entre os participantes da pesquisa. Também são apresentadas estatísticas referentes ao impacto dos spams identificados na reputação online dos produtos.

Ao final, propõe-se o desenvolvimento de um modelo computacional utilizando técnicas de Aprendizagem de Máquina, a fim de classificar automaticamente opiniões legítimas e possíveis spams.

No capítulo 2, o tema spams de opinião é aprofundado, elencando seus tipos e as principais características de pesquisas já publicadas, as quais visaram identificá-los de maneira automática. Em seguida, o capítulo 3 descreve como os conjuntos de opiniões foram criados e anotados por outros autores, visando o desenvolvimento de uma metodologia própria, apresentada no capítulo 4. O capítulo 5 discute os resultados e confiabilidade da anotação, para que no sexto capítulo, seja proposto o estudo preliminar de um classificador de opiniões. Por fim, são apresentadas conclusões sobre a pesquisa.

2 OPINIÕES NO AMBIENTE ONLINE

O que outras pessoas pensam sempre foi importante durante o processo de tomada de decisão. Muito antes da *World Wide* Web (WWW), era comum pedir recomendações para amigos sobre prestadores de serviços e até em quem iriam votar nas eleições. Mas a Web tornou possível encontrar opiniões e experiências de indivíduos dos quais nunca ouvimos falar (PANG; LEE, 2008).

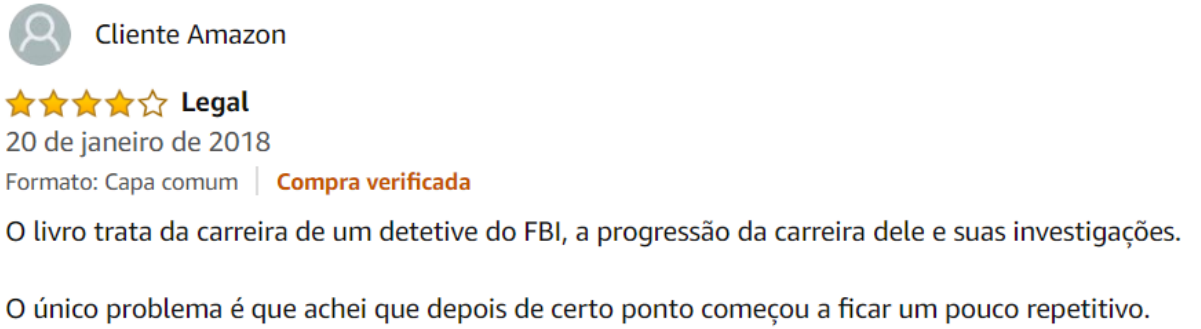
As redes sociais, devido à evolução da Internet e do crescimento no uso de *smartphones*, impulsionaram a geração e consulta de opiniões sobre produtos, serviços, estabelecimentos, pessoas e fatos. Consumidores passaram a analisá-las antes de comprar um produto (CARDOSO; ALMEIDA, 2017). Liu (2012, p. 113, tradução nossa) corrobora nesse sentido, afirmando que “opiniões de mídias sociais estão sendo crescentemente usadas por indivíduos e organizações para tomar decisões de compra”. Estudos de Murphy (2018) apontam que, no Reino Unido, 80% das pessoas entre 18 e 34 anos de idade já escreveram suas opiniões na Internet.

De acordo com o Dicionário da Língua Portuguesa Michaelis (2019, on-line), opinião pode ser entendida como um “modo de pensar, de julgar, de ver” ou ainda, “um parecer emitido sobre determinado assunto em que muito se refletiu e deliberou”. Já Liu (2012), apresenta uma explicação mais técnica. Segundo o autor, opiniões são compostas basicamente de dois elementos-chave: o alvo (*g*) e o sentimento sobre o alvo (*s*). O componente *g* pode ser uma entidade ou uma característica dela, sobre a qual uma opinião foi expressa. Já *s* é um sentimento positivo, negativo ou neutro, que geralmente é descrito com um índice numérico, representando sua intensidade.

Com o auxílio da Figura 1, é possível identificar tais elementos em uma opinião extraída do site Amazon[1], empresa de mercado eletrônico, considerada uma das mais valiosas do mundo (BBC, 2019). O índice *s*, também chamado de *rating,* é representado pelo número de estrelas. Quanto maior, mais satisfeito o usuário está.

[1] Site da Amazon disponível em: www.amazon.com.br

Figura 1 – Opinião sobre um produto à venda na Amazon



Fonte: adaptado de Amazon (2019)

A Figura 1 ainda apresenta outros dois aspectos descritos por Liu (2012) como compositores da opinião. A saber: o detentor (*h*), geralmente o usuário que a postou e a data de postagem (*t*). Logo, uma opinião é formada pelo quádruplo *g*, *s*, *h*, *t* (LIU, 2012).

Hu e Liu (2004) argumentam que, com o propósito de aumentar a satisfação dos usuários, vendedores permitiram que seus clientes avaliassem produtos que compraram. Logo, é reconhecido que o conteúdo criado por usuários, do inglês *user-generated content*, como avaliações sobre produtos, contém informações valiosas, as quais podem ser utilizadas em diversas aplicações. Trabalhos existentes focam principalmente na sumarização de opiniões (JINDAL; LIU, 2008), a qual é a atividade de extrair, sintetizar e visualizar as informações mais relevantes (CONDORI; PARDO, 2017). A sumarização das avaliações é importante, dado o grande número de *reviews*. Consumidores em potencial podem ter uma visão distorcida da mercadoria caso não consigam ler todos os textos disponíveis (HU; LIU, 2004).

Apesar disso, Jindal e Liu (2008) explicam que pouco se sabe sobre as características das avaliações e o comportamento dos usuários que as postam. Os autores buscaram estudar a veracidade das opiniões em *reviews*. “Dado que não há controle de qualidade, qualquer um pode escrever qualquer coisa na Internet. Isso resulta em muitas avaliações de baixa qualidade e pior, até spams em *reviews*” (JINDAL; LIU, 2008, p. 1, tradução nossa).

Mukherjee, Liu e Glance (2012) explicam que, se uma pessoa quiser comprar um produto, irá ler suas avaliações. Se a maioria for positiva, tenderá a comprá-lo, mas se grande parte for negativa, provavelmente o consumidor escolherá outra mercadoria. Esse cenário, de acordo com Liu (2012), proporciona fortes incentivos

para a atividade conhecida como spams de opinião, que se refere a atividade humana de enganar leitores ao escrever *reviews* injustas sobre determinados alvos. Os indivíduos que a praticam, por sua vez, são conhecidos como *spammers*.

Spams de opinião são extremamente diferentes dos encontrados em e-mails e na Web, dois dos tipos mais estudados (LIU, 2012). O primeiro refere-se a links para outros sites indesejados. Enquanto o segundo é caracterizado pela inserção de palavras populares em sites, a fim de que motores de pesquisa os marquem como relevantes (LIU, 2012). A popularidade desse termo relacionado com opiniões começou a crescer a partir de 2008, através da publicação de Jindal e Liu (2008), os primeiros a empregar esforços nessa área. Contudo, spams ainda estão diretamente vinculados com e-mails. Ao consultar o significado dessa expressão no Dicionário da Língua Portuguesa Michaelis (2019, on-line), por exemplo, tem-se como definição: “mensagem eletrônica indesejada que chega aos computadores”.

Pelo fato de que “opiniões positivas geralmente significam lucro e fama para negócios e indivíduos” (LIU, 2012, p. 113, tradução nossa), comerciantes criam avaliações falsas, as quais apresentam dois efeitos prejudiciais para os consumidores. Primeiramente, elas induzem compradores a tomar decisões ruins. Em segundo lugar, a confiança em avaliações online cai, já que para o leitor, pareceu ser uma boa ideia adquirir o item pela quantidade de elogios recebidos (SANDULESCU; ESTER, 2015). Portanto, é essencial que spams em *reviews* sejam detectados a fim de garantir que as mídias sociais continuem sendo uma fonte confiável de opiniões públicas, ao invés de ser infestada com mentiras e fraudes (LIU, 2012).

Após explicações iniciais, é necessário aprofundar-se nos tipos de spams de opinião identificados pela literatura, a fim de esclarecer suas características, bem como elucidar de que maneiras eles podem estar inseridos no ambiente online.

2.1 TIPOS DE SPAMS DE OPINIÃO

Jindal e Liu (2008) propuseram a divisão dos spams de opinião em três tipos, devido à sua dificuldade de detecção. São eles:

- **Tipo 1 (*reviews* falsas)**: Não são escritas com base na real experiência do avaliador, mas sim com um motivo oculto. R*eviews* positivas para produtos

ou serviços que não as merecem, bem como opiniões negativas que servem somente para difamar a reputação do alvo;

- **Tipo 2 (*reviews* sobre marcas)**: Não comentam sobre produtos ou serviços específicos, somente nas marcas e fabricantes. Embora possam ser genuínas, são consideradas spam por não abordarem as características da entidade sendo avaliada; e
- **Tipo 3 (não-*reviews*)**: Propagandas e outros textos irrelevantes, os quais não possuem opiniões.

Jindal e Liu (2008) afirmam que os tipos 2 e 3 de spam são mais fáceis de serem identificados automaticamente, utilizando-se de exemplos etiquetados como spam e não-spam. E mesmo que não sejam detectados, não chegam a representar grandes problemas, dado que humanos podem encontrá-las facilmente durante a leitura (LIU, 2012). Um exemplo do tipo 2, segundo Liu (2012, p. 114, tradução nossa) poderia ser “Eu odeio a HP. Nunca compro produtos deles”. Ao se referir à empresa HP (Hewlett-Packard), o postador não aborda as características do produto, sendo irrelevante para análises da qualidade da mercadoria.

A Figura 2 apresenta uma não-*review* (tipo 3). Em nenhum momento o usuário menciona o produto. Ele apenas faz críticas à falta de disponibilidade do mesmo na loja, não descrevendo uma opinião de fato. Um fator que chama a atenção é o *rating* da avaliação. Das dez estrelas possíveis, foi dado justamente a nota máxima, mesmo que o usuário não tenha experimentado o produto.

Figura 2 – Exemplo de uma não-*review*

Fonte: Buscapé (2019)

Já o primeiro tipo, no mais extremo dos casos, pode tornar impossível sua identificação através da anotação manual pela leitura das *reviews*, visto que elas podem ser cuidadosamente escritas pelos *spammers* (LIU, 2012). Por isso,

complementa Liu (2012, p. 113, tradução nossa), “é difícil encontrar dados sobre spams de opinião que ajudem a projetar e avaliar algoritmos de detecção”.

Reviews falsas, entretanto, não são igualmente prejudiciais (LIU, 2012). Assumindo que é sabido a real qualidade do produto, o Quadro 1 demonstra que há seis diferentes relações entre as avaliações falsas e a qualidade do produto. Nas regiões A, C e E, o objetivo é promover o produto. Já em B, D e F, é de atacar sua reputação.

Liu (2012) explica que as *reviews* das regiões A e F não são deveras prejudiciais, já que a opinião é compatível com a qualidade do produto. Porém, *spammers* ainda possuem segundas intenções ao postá-las. Já as demais regiões B, C, D e F (em negrito), são mais nocivas e devem ser o foco de algoritmos para detecção automática de *reviews* falsas.

Quadro 1 – Reviews falsas x qualidade do produto

Qualidade do Produto	*Review* falsa positiva	*Review* falsa negativa
Boa	A	**B**
Média	**C**	**D**
Ruim	**E**	F

Fonte: adaptado de Liu (2012)

Liu (2012) explica que *spammers*, apesar de agirem predominantemente sozinhos, também podem fazer parte de um grupo, mesmo que não saibam. Mukherjee, Liu e Glance (2012) afirmam que grupos de *spammers* possuem mais mão-de-obra humana e assim, cada membro pode não parecer estar se comportando anormalmente. Em resumo, há duas subclasses de *group spamming* definidas por Mukherjee, Liu e Glance (2012):

- Um grupo de pessoas que age em conjunto, sendo que os integrantes podem ou não se conhecer; e
- Uma única pessoa que cria múltiplas contas em *websites*, simulando um grupo de indivíduos.

Ao contrário das práticas individuais, onde *spammers* trabalham sozinhos utilizando apenas uma conta, *spamming* em grupo tende a ser mais prejudicial, visto que se torna mais fácil controlar o sentimento sobre algum produto, enganando consumidores em potencial (MUKHERJEE; LIU; GLANCE, 2012).

Para a detecção desses tipos de spams, Liu (2012) argumenta que há três tipos de dados que podem ser analisados sobre a *review* ou o produto-alvo:

- **Conteúdo da *review*:** o real texto da opinião, nos quais é possível extrair características linguísticas;
- **Metadados da *review*:** dados sobre as *reviews*, como código do usuário, tempo utilizado por ele para escrever a opinião, geolocalização do postador, entre tantas outras possibilidades. Tais informações podem ser usadas para detectar anomalias comportamentais dos usuários; e
- **Informações do produto:** dados sobre a entidade sendo avaliada, como volume de vendas *versus* reputação do mesmo no ambiente online.

Após a identificação dos diferentes tipos de spams de opinião e como, em teoria, informações podem ser levantadas para identificá-los, é possível nos aprofundarmos em estudos recentes de empresas e pesquisadores sobre tentativas de solucionar o problema das *reviews* falsas ou suspeitas, através de algoritmos computacionais.

2.2 DETECÇÃO DE SPAMS EM OPINIÕES

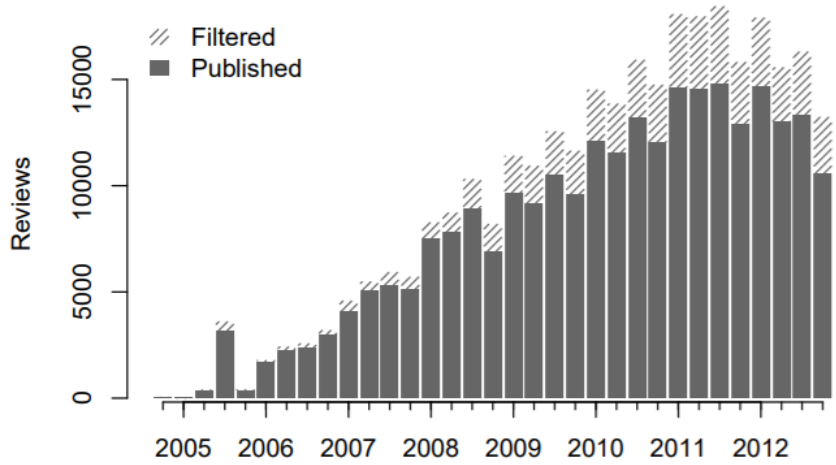
A fim de elucidar a dificuldade em identificar os diferentes tipos de spams em opiniões, é apresentado o caso do Yelp[2], um site de *reviews* online em grande escala, o qual tenta filtrar avaliações falsas, ou no mínimo, suspeitas (MUKHERJEE et al., 2013). Estudos de Mukherjee et al. (2013) comprovaram que o filtro aplicado pela empresa pode ser considerado como razoável e confiável. A Figura 3 demonstra o crescimento das *reviews* em restaurantes de uma cidade dos Estados Unidos filtradas pelo algoritmo. De acordo com Luca e Zervas (2016), 16% das *reviews* do Yelp sobre este serviço são spams.

Observa-se no gráfico que o crescimento das avaliações consideradas como falsas, representado pelas barras tracejadas, acompanham proporcionalmente o número total de *reviews*. O aumento no número de opiniões filtradas pelo Yelp, entretanto, ainda recebe constantes críticas. Dentre as principais, segundo Raleigh (2018), estão donos de estabelecimentos descontentes com o algoritmo utilizado pela

[2] Site do Yelp disponível em: www.yelp.com

companhia, pois classifica *reviews* de cinco estrelas como não recomendadas, ou oculta opiniões de usuários que não são *spammers*. Apesar disso, o Yelp não disponibiliza explicações sobre como seu algoritmo de filtragem funciona (RALEIGH, 2018).

Figura 3 – Número de reviews filtradas pelo Yelp



Fonte: Luca e Zervas (2016)

Outros exemplos, como o da Amazon, também demonstram a vontade das organizações em aumentar a credibilidade de seus sistemas. Cardoso e Almeida (2017) explicam que o Amazon *Verified Purchase* (AVP) é uma garantia de que o autor de determinada avaliação efetuou a compra do produto. “Este recurso tenta minimizar uma possível desconfiança em *reviews* online e parte do esforço da Amazon em combater este tipo de atividade fraudulenta” (CARDOSO; ALMEIDA, 2017). Em 2015, a companhia entrou com ações judiciais contra sites que vendiam avaliações falsas de consumidores para comerciantes da Amazon (BISHOP, 2015).

Desde os primeiros estudos aplicados na área, Sandulescu e Ester (2015) comentam que a maioria das pesquisas seguiram duas direções: análises de características comportamentais e textuais. Assim como os tipos de dados apontados por Liu (2012) citados na seção 2.1, Sandulescu e Ester (2015) elucidam que do ponto de vista comportamental, podem ser utilizadas informações como a data da *review*, seu *rating* médio, o endereço IP de onde a mesma foi postada e assim por diante. Já do lado textual, sua análise se baseia em extrair pistas do conteúdo da avaliação, desde padrões de texto até a frequência das palavras.

Sandulesco e Ester (2015) e Ott et al. (2011), por exemplo, deram ênfase na análise textual. Os primeiros propuseram um estudo de similaridade semântica, a fim de identificar *spammers* que escrevem *reviews* falsas com diferentes nomes. De acordo com os autores, a análise de padrões comportamentais é útil apenas em usuários de elite, ou seja, que postam constantemente. “Para avaliadores de uma só vez, o texto das opiniões deve ser explorado” (SANDULESCU; ESTER, 2015). Já Ott et al. (2011) elaboraram modelos para identificar spams, introduzindo análises de emoções, através de atributos psicolinguísticos.

Referente à abordagem comportamental, Li et al. (2015) basearam-se em dados espaciais e temporais das reviews sobre estabelecimentos, a fim de identificar em quais períodos spams são mais comumente postados, bem como verificar se os usuários realmente estiveram nos locais avaliados recentemente.

Lim et al. (2010) também seguiram essa vertente, analisando *ratings* de *reviews* que destoam do padrão de estrelas do produto. Segundos os autores, para promover ou descreditar produtos, os avaliadores inclinam-se a dar *ratings* que diferem da média. Outro ponto identificado por Lim et al. (2010), é que *spammers* tendem a atacar produtos ou grupos de produtos específicos, principalmente logo após ficarem disponíveis para avaliações nos *websites*. Logo, características temporais também são importantes para a tarefa de mitigar o impacto de *reviews* não genuínas.

Apesar da distinção entre as duas vertentes, a maioria dos estudos realizados as combinaram com o objetivo de maximizar a eficácia dos resultados, como demonstrado por Jindal e Liu (2008), Yuan et al. (2016), Mukherjee e Venkataraman (2014), Mukherjee, Liu e Glance (2012) e Cardoso e Almeida (2017).

Jindal e Liu (2008) optaram por analisar opiniões com textos quase ou totalmente duplicados. As *reviews* que possuíam cópias no conjunto de avaliações utilizadas, foram rotuladas como spam, enquanto as outras, como não-spam. A partir dessa premissa, os autores utilizaram atributos como o *rating* e quantidade de usuários que acharam aquela opinião útil para auxílio na montagem do modelo de classificação automática.

Yuan et al. (2016) estudaram variações no número de opiniões em função do período de postagem das *reviews*. Anomalias em padrões identificados ao longo do tempo podem ser indicativos de *spamming* (YUAN et al., 2016). A quantidade de

reviews, rating da *review* comparada com o *rating* médio do alvo, além de análises de sentimento presentes na opinião, ou seja, com base no texto, identificar a polaridade (positiva ou negativa) do sentimento que o usuário teve com o alvo, foram outros atributos avaliados por Yuan et al. (2016).

Já Mukherjee e Venkataraman (2014), abordaram a similaridade dos textos, mas também o desvio padrão do *rating* das *reviews* e período em que a opinião foi postada, combinando assim, as propostas de Jindal e Liu (2008) e Lim et al. (2010).

Todos os autores citados acima propuseram o desenvolvimento de algoritmos de detecção automática de spams, utilizando técnicas de Aprendizagem de Máquina, do inglês, *Machine Learning* (ML). Para isso, recorreram ao uso de conjuntos de opiniões, sobre os quais realizaram análises textuais e/ou comportamentais, para então criar modelos computacionais que classifiquem novas *reviews*.

Embora importantes, programas de ML necessitam, primeiramente, de exemplos previamente anotados, ou seja, rotulados como spam ou não-spam, para que então, seja possível identificar padrões sobre os dados.

Ao longo de toda a obra de Liu (2012), a palavra corpus é utilizada para se referir ao agrupamento de palavras, textos e metadados sobre os textos. Por corpus, o Dicionário da Língua Portuguesa Michaelis (2019) entende como sendo o “conjunto de documentos e informações sobre determinado assunto”, ou ainda “conjunto de enunciados de uma língua, que é utilizado como material para análise linguística”. O conjunto de corpus é conhecido por corpora (MANNING; SCHÜTZE, 1999).

A Figura 4 apresenta um extrato em *Extensible Markup Language* (XML) retirado do corpus CETENFolha[3], o qual contém cerca de 24 milhões de palavras do português brasileiro, criado a partir de textos do jornal Folha de São Paulo. Ele foi concebido com o intuito de ser matéria-prima para programas que processem a língua portuguesa (LINGUATECA, 2002). Nas *tags* do extrato, encontram-se informações como título, autores e parágrafos do texto, os quais podem ser usados como base para análises linguísticas.

[3] Site do CETENFolha disponível em: www.linguateca.pt/cetenfolha/index_info.html

Figura 4 – Extrato do CETENFolha

```
<ext id="165572" cad="Ilustrada" sec="nd" sem="94b">
    <s><t>Batata será matéria-prima no futuro</t></s>
    <s><a>Das agências internacionais</a></s>
    <p><s>Os poloneses acreditam no futuro da batata,
          cuja produção no país já supera 40 milhões de toneladas ao ano.</s></p>
    <p><s>Segundo os pesquisadores, com o desenvolvimento de novas variedades,
          como as  resistentes à seca e à umidade excessiva, ela se tornará no
          século 21 um dos principais alimentos do homem, além de grande fornecedor
          de matéria-prima para fármacos e cosméticos.</s></p>
</ext>
```

Fonte: adaptado de Linguateca (2002)

Utilizando-se de corpus como esse, diversos autores desenvolveram algoritmos de aprendizagem automática. O próximo capítulo explica que a anotação de corpus, embora custosa, representa um passo essencial para o desenvolvimento de algoritmos de classificação.

3 CRIAÇÃO E ANOTAÇÃO DE CORPUS

Fundamentando-se em uma explicação sobre spams em e-mails, Shalev-Shwartz e Ben-David (2014) introduzem o conceito de *Machine Learning*. De acordo com os autores, para que uma máquina classifique automaticamente e-mails como spam ou não, é necessário disponibilizar um conjunto de exemplos já anotados, que podem ser fornecidos por um usuário humano. A partir de então, a máquina poderá extrair características dos dados de treinamento, utilizando-as como base para rotular novos e-mails (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Russel e Norvig (2013, p. 605) definem ML como, “a partir de uma coleção de pares de entrada e saída, aprender uma função que prevê a saída para novas entradas”. Programar um agente para aprender com suas próprias experiências é útil quando os projetistas não têm como antecipar todas as situações possíveis, nem mudanças ao longo do tempo. Por isso, o programa deve ser capaz de se adaptar de acordo com as condições, propondo soluções adequadas (RUSSEL; NORVIG, 2013).

Mesmo após uma breve explanação, já é possível perceber a necessidade do conjunto de exemplos em ML. Portanto, antes de discutir conceitos e técnicas de Aprendizado de Máquina aplicadas em spams de opinião, faz-se necessário o entendimento sobre como obter informações suficientes para treinamento, bem como garantir a confiabilidade dos exemplos anotados. De acordo com Shalev-Shwartz e Ben-David (2014, p. 20, tradução nossa)

> Essa questão é crucial para o desenvolvimento de aprendizagens automáticas. Enquanto humanos podem se basear no senso comum para filtrar conclusões de aprendizado sem significado, uma vez que exportamos a tarefa de aprender para uma máquina, precisamos criar princípios bem definidos que protegerão o programa de chegar em conclusões sem sentido e inúteis. O desenvolvimento de tais princípios é o objetivo central da teoria do aprendizado de máquina.

A seção 3.1 apresenta corpora de opiniões já criados por outros autores. O intuito é identificar práticas comuns utilizadas para a montagem e anotação de dados de treinamento. Após isso, são descritos métodos que buscam garantir a confiabilidade dos exemplos etiquetados por seres humanos.

3.1 CORPORA PESQUISADOS E EXEMPLOS DE ANOTAÇÕES

Dada a dificuldade em encontrar corpora com opiniões reais anotadas como spams e não-spams, discutida na seção 2.1, algumas pesquisas buscaram criar bases de dados de diferentes maneiras, aplicando métodos distintos de anotação.

Jindal e Liu (2008) analisaram o conteúdo de 5,8 milhões de *reviews* extraídas do site da Amazon. Opiniões quase ou totalmente duplicadas, ou seja, textos que possuíam alto grau de similaridade foram classificados como spams. Essa anotação foi utilizada, posteriormente, por algoritmos de classificação, visando a identificação de dados comportamentais das *reviews,* importantes para a correta classificação das opiniões. Mukherjee, Liu e Glance (2012) realizaram pesquisas sobre identificação de grupos de *spammers* sobre o mesmo corpus. Porém, para rotular possíveis grupos, especialistas na área apontaram os candidatos. Em seguida, os pesquisadores utilizaram o coeficiente Kappa de Fleiss, com o propósito de medir o grau de concordância entre os peritos.

O coeficiente Kappa de Fleiss será aprofundado na seção 3.2, mas é importante elucidar que métricas como essa geram índices que representam o grau de concordância em que diferentes jurados chegaram sobre a etiquetação de unidades do corpus (COHEN, 1960).

Ott et al. (2011) recolheram 400 *reviews* falsas, intencionalmente fabricadas por usuários do *Amazon Mechanical Turk* (AMT). O AMT é um serviço de *crowdsourcing,* no qual é possível criar tarefas para qualquer pessoa inscrita na plataforma executá-la. Avaliações verdadeiras foram retiradas do TripAdvisor, outro site de opiniões sobre estabelecimentos que contém *user-generated content* (OTT et al., 2011). Com o corpus formado, extraíram características textuais, a fim de desenvolver um modelo de ML capaz de anotá-las automaticamente.

Baseando-se no algoritmo do Yelp, introduzido na seção 2.2, Mukherjee et al. (2013), Sandulescu e Ester (2015) e Yao et al. (2018) o utilizaram com o objetivo de propor modelos de identificação de opiniões falsas. Para treinamento, as *reviews* não recomendadas pelo programa da empresa foram consideradas como spam, enquanto as demais foi assumido que eram verdadeiras. O Quadro 2 demonstra os atributos do corpus apresentado por Yao et al. (2018). A propriedade “recomendada” é responsável por conter a etiquetação das *reviews*.

Quadro 2 – Atributos sobre uma review do Yelp

Categoria	Descrição
Data	Data formatada
Número de amigos	Número de amigos do usuário
Tem foto de perfil	Verdadeiro ou falso para foto do perfil do usuário
Local	Cidade e Estado do usuário
Número de fotos	Quantidade de fotos tiradas pelo usuário
Rating	Rating de 1 a 5 dado pelo usuário
ID do Restaurante	Código do restaurante sobre o qual foi postada a *review*
Número de reviews	Número de *reviews* feitas pelo usuário
Texto	Texto da *review*
Nome do usuário	Primeiro nome e última inicial do usuário
Recomendada	0 para falso, 1 para verdadeiro

Fonte: adaptado de Yao et al. (2018)

O Yelp promove desafios sobre dados abertos, os quais convidam o público para fazer descobertas a partir de suas informações. Entretanto, a base de dados original do Yelp não contempla as *reviews* não recomendadas, sendo necessário extraí-las diretamente do site (YAO et al., 2018).

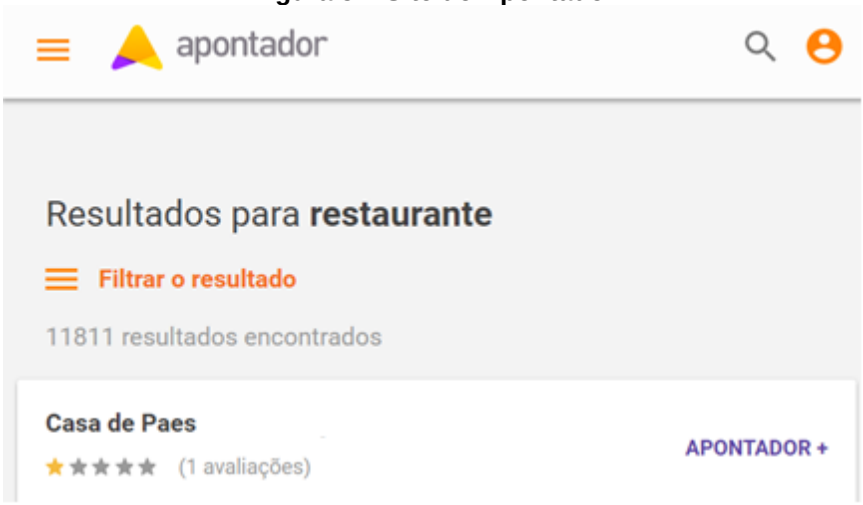
Li et al. (2015) também incluíram outro corpus nessa lista. Os autores analisaram mais de 6 milhões de opiniões disponibilizadas pelo Dianping, um site hospedeiro de *reviews* sobre restaurantes de Xangai (LI et al., 2015). Assim como o Yelp, o Dianping também possui um filtro de avaliações falsas. Semelhante ao executado em Mukherjee et al. (2013), *reviews* filtradas pelo algoritmo da empresa foram considerados como spam.

Demais trabalhos também foram responsáveis por introduzir outras bases de dados de opiniões, como Zhang et al. (2012) e Li et al. (2011), embora seus métodos de extração e anotação sejam semelhante aos já citados.

É possível perceber que os avanços na área de spams de opinião se dão, na sua grande maioria, associados à língua inglesa. Poucos esforços foram empreendidos para o português. Até o momento, foi encontrado apenas o trabalho de Costa, Benevenuto e Merschmann (2013) que aborda o problema nesse idioma. Os autores avaliaram *reviews* extraídas do Apontador[1] (ilustrado na Figura 5), uma Rede Social Baseada em Localização (RSBL).

[1] Site do Apontador disponível em: www.apontador.com.br

Figura 5 – Site do Apontador

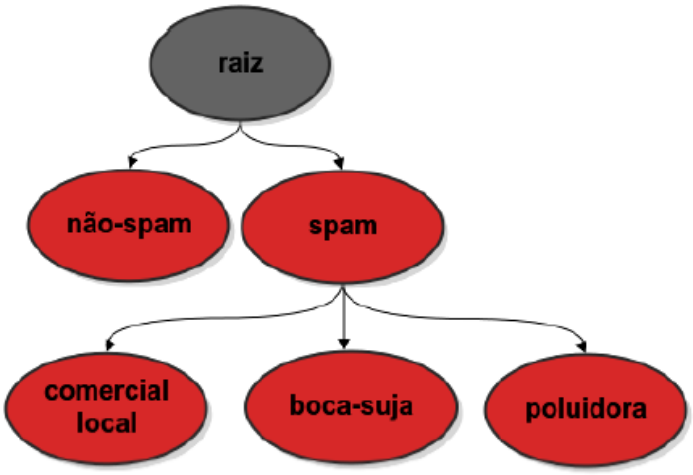


Fonte: Apontador (2019)

Cada vez mais, as RSBL veem atraindo novos adeptos, pois permitem que usuários compartilhem suas localizações via geolocalização com *smartphones* de amigos e pesquisem por lugares de interesse. Além disso, o Apontador permite a postagem de avaliações sobre locais existentes (COSTA; BENEVENUTO; MERSCHMANN, 2013).

Costa, Benevenuto e Merschmann (2013) montaram seu corpus inicial com 7.336 opiniões disponibilizadas e anotadas pelos próprios moderadores do site como 50% spam e os outros 50% como não-spam. A Figura 6 demonstra os tipos de spam encontrados. Em termos gerais, eles foram divididos em três classes: Comercial Local, Poluidora e Bocas-sujas.

Figura 6 – Tipos de spam encontrados em *reviews* do Apontador



Fonte: Costa, Benevenuto e Merschmann (2013)

A classe Comercial Local se refere a propagandas sobre o local-alvo. Poluidoras são caracterizadas por seu conteúdo irrelevante ou não relacionado ao local. Enquanto o tipo Bocas-sujas contém comentários agressivos e palavras de baixo calão (COSTA; BENEVENUTO; MERSCHMANN, 2013).

A fim de aumentar a confiabilidade na anotação humana, um grupo de voluntários revisou a etiquetação de todas as *reviews* spam. Apenas em 3,5% das 3.668 opiniões, os voluntários discordaram dos moderadores do Apontador. Isso demonstra um alto grau de confiabilidade na anotação (COSTA; BENEVENUTO; MERSCHMANN, 2013). A discordância entre os administradores do Apontador e os voluntários ocasionaram no descarte de 130 avaliações spam. Assim, o corpus final balanceado, utilizado para o desenvolvimento de classificadores em ML totalizou 7.076 *reviews*, divididas igualmente entre as duas classificações possíveis.

Nos trabalhos acima, é perceptível a preocupação dos pesquisadores com a confiabilidade das etiquetações feitas por humanos, visto que serão a base para o desenvolvimento de algoritmos de detecção de spams de opinião. Foi possível averiguar que o método mais comum para garantir confiança nos exemplos era medir o nível de acordo entre os anotadores. Por isso, a seção 3.2 exemplifica maneiras de calculá-lo.

3.2 CONCORDÂNCIA ENTRE ANOTADORES

Estudos que medem a concordância entre dois ou mais observadores devem incluir uma estatística que aceita o fato de que eles irão concordar ou discordar entre si (VIERA; GARRETT, 2005). Ao descrever a literatura em testes de diagnósticos médicos, Viera e Garrett (2005) explicam que tais testes se baseiam em certo grau de subjetividade dos observadores, aqueles que realmente interpretam os resultados. Portanto, se eles não concordarem nas suas interpretações, os resultados serão de pouco uso (VIERA; GARRETT, 2005).

Cohen (1960) descreve que, em atividades nas quais jurados categorizam uma amostra, é importante determinar a extensão até a qual os julgamentos são reprodutíveis e confiáveis, com o propósito de “determinar o grau, significância e estabilidade na concordância entre os mesmos” (COHEN, 1960, p.37-38, tradução nossa). Dosciatti, Ferreira e Paraiso (2015, p. 125) complementam, afirmando que

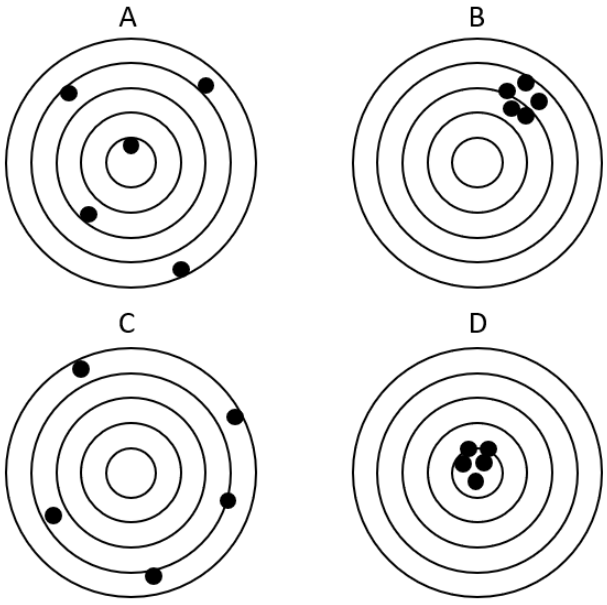
> O percentual de casos em que dois anotadores concordam em relação à classificação de um conjunto de textos com um dado número de categorias é a forma mais simples de se atribuir confiabilidade a um processo de anotação de textos realizado em determinado corpus.

Dois conceitos importantes devem ser entendidos *a priori*: precisão e acurácia. Embora semelhantes, ambos possuem características distintas, essenciais para averiguarmos os reais objetivos da métrica para calcular a concordância entre anotadores.

Viera e Garrett (2005) descrevem esses conceitos através da analogia com alvos, ilustrada na Figura 7. Se o centro do alvo for atingido por um tiro, isso representa acurácia, mas se vários tiros se agruparem em um ponto do alvo, tem-se precisão.

O alvo A refere-se à acurácia, visto que um dos tiros encontrou o centro, apesar de não haver precisão, pois os demais pontos estão descentralizados. Em B, é possível verificar a precisão, mas não acurácia. Em C, nem um nem outro. Já em D, todos os pontos encontram-se agrupados (precisão) e no centro do alvo (acurácia) (VIERA; GARRETT, 2005).

Figura 7 – Precisão x Acurácia



Fonte: adaptado de Viera e Garrett (2005)

“A falta de precisão em A e C pode ser ao acaso, que nesse cenário, o tiro no centro do alvo A foi sorte. Em B e D, o agrupamento não é provável devido ao acaso” (VIERA; GARRETT, 2005).

A concordância entre anotadores portanto, está diretamente associada à precisão (VIERA; GARRETT, 2005). Viera e Garret (2005) afirmam que esse processo garante boa confiança, pois os exemplos obtidos não são baseados em achismos. Embora não signifique que haja acurácia, pois há a possibilidade de todos os anotadores terem errado suas observações. McHugh (2012) elenca algumas das estatísticas que visam medir o grau de concordância, incluindo o percentual simples, Kappa de Cohen, Kappa de Fleiss e Alpha de Krippendorff.

Da lista proposta por McHugh (2012), foram omitidos os modelos que verificam a correlação entre variáveis discretas, visto que spams de opinião normalmente se classificam em duas classes: spam e não spam (LIU, 2012), portanto são variáveis categóricas. Por muito tempo, a concordância entre avaliadores foi medida pelo percentual de concordância (MCHUGH, 2012). A seguir, são demonstrados alguns cálculos para verificar o nível de acordo em tarefas de etiquetação. Primeiramente com o percentual de concordância, depois aprofundando-se no coeficiente Kappa e suas variações.

3.2.1 Percentual de concordância

De acordo com McHugh (2012), este cálculo pode ser apresentado na forma de uma matriz, na qual as colunas representam diferentes anotadores, enquanto as linhas são unidades sobre as quais as rotulações são coletadas. As unidades podem representar, por exemplo, opiniões sobre produtos. A Tabela 1 ajuda a elucidar tal processo. O pesquisador simplesmente calcula o percentual de acordo para cada linha e em seguida, encontra a média dos resultados.

Tabela 1 – Percentual de concordância entre múltiplos anotadores

	Anotadores					
Unidades	Marcos	Susana	Tomás	Ana	Joice	Concordância
1	1	1	1	1	1	1,00
2	1	1	1	1	1	1,00
3	0	1	1	1	1	0,80
4	0	0	0	0	0	1,00
5	1	1	0	0	1	0,60
Confiança de anotação no estudo						0,88 (88%)

Fonte: adaptado de McHugh (2012)

No cenário acima, a confiança é de 88%. Isso significa que 12% dos dados captados são errôneos (MCHUGH, 2012). Contanto que a quantidade de valores possíveis de anotação para cada unidade permaneça em dois, o cálculo se mantém simples (MCHUGH, 2012). Para McHugh (2012), esse método permite identificar se um avaliador em particular registra valores diferentes dos demais, além de encontrar variáveis problemáticas, como é o caso da nº 5, já que ela contém apenas 60% de concordância.

Entretanto, há uma séria discussão sobre essa métrica. Segundo Cohen (1960, p. 38, tradução nossa), “não há critério para exatidão nos julgamentos”, pois leva-se em consideração que a maioria está correta. Logo, a minoria está incorreta em suas anotações (MCHUGH, 2012).

Por isso, Cohen (1960) sugeriu que parte dos julgamentos de pelo menos algumas unidades, nenhum dos avaliadores estavam certos da sua observação, tendo opiniões aleatórias. Levando isso em consideração, é possível que o consenso encontrado nas variáveis pode ser falso (MCHUGH, 2012). A próxima seção descreve o que Cohen (1960) propôs para resolver este problema.

3.2.2 O coeficiente Kappa

Segundo Freitas e Vieira (2013), o coeficiente Kappa é uma métrica que avalia o nível de concordância em tarefas de classificação. Refere-se à precisão do nível de concordância entre os observadores apenas sobre problemas de categorização nominal (COHEN, 1960), aferindo confiabilidade (VIERA; GARRETT, 2005).

Cohen (1960) explica que a forma mais primitiva de se avaliar o grau de concordância entre jurados é simplesmente somar o número de vezes em que eles concordaram. Porém, em certas vezes, o acordo pode se dar justamente ao acaso. Com a finalidade de resolver a “inadequação desta solução” (COHEN, 1960, p. 38, tradução nossa), o autor propôs um novo método, conhecido como Kappa.

“O coeficiente Kappa leva em conta no cálculo, a proporção de concordância que é devido ao acaso e por isso, é bastante utilizado para medir a concordância entre anotadores em corpora” (DOSCIATTI; FERREIRA; PARAISO, 2015, p. 125). O índice calculado pode variar entre 1 e menor que 0. Kappa igual a 1 indica acordo perfeito,

enquanto Kappa mais próximo de 0 indica concordância equivalente ao acaso (VIERA; GARRETT, 2005).

De acordo com Cohen (1960) há algumas premissas que devem ser levadas em conta para a aplicação do cálculo:

- As unidades anotadas são independentes;
- As categorias nominais são independentes umas das outras, mutualmente exclusivas e exaustivas, ou seja, uma unidade não pode assumir duas classificações ao mesmo tempo, muito menos não possuir nenhuma;
- Os jurados operam independentemente;
- Os jurados são declarados competentes para fazer julgamentos *a priori*;
- Não há restrições quanto à distribuição de julgamentos para qualquer jurado.

O método proposto por Cohen (1960) para o cálculo do Kappa entretanto, não suporta mais de dois anotadores julgando o mesmo conjunto de unidades. Com o auxílio das Tabelas 2 e 3, é possível explicar com clareza tanto o método tradicional, quanto o Kappa de Cohen. Nela, dois jurados classificam o mesmo conjunto de exemplos em duas categorias possíveis. Em *a* e *d* tem-se a soma de vezes em que os jurados acordaram, enquanto em *b* e *c*, as vezes em que eles discordaram.

Tabela 2 – Modelo de Matriz de concordância entre dois jurados

		Jurado 1		
		Categoria 1	Categoria 2	Total
Jurado 2	Categoria 1	a	b	m_1
	Categoria 2	c	d	m_0
	Total	n_1	n_0	n

Fonte: adaptado de Viera e Garrett (2005)

Supondo que tenham sido apresentados 100 textos de opiniões para serem classificados como spam (sim e não) por dois jurados, a Tabela 3 apresenta dados fictícios. Sobre o cálculo do coeficiente, Viera e Garrett (2005, p. 361, tradução nossa) afirmam que ele “é baseado na diferença entre o quanto de concordância está realmente presente (concordância observada) comparada com o quanto de concordância é esperada por acaso (concordância esperada)”.

Tabela 3 – Exemplo de Matriz de concordância

		Jurado 1		
		Sim	Não	Total
Jurado 2	Sim	15	5	20
	Não	10	70	80
	Total	25	75	100

Fonte: adaptado de Viera e Garrett (2005)

Para calcular a concordância observada (p_o), utiliza-se a fórmula:

$$p_o = \frac{a + d}{n}$$

No exemplo da Tabela 3, p_o terá o valor de 0,85, visto que a soma entre *a* (15) e *d* (70) será dividida pelo total *n* (100). No modelo tradicional, 85% já seria considerado como percentual de concordância definitivo, sem levar em consideração as situações onde os acordos sejam atingidos por acaso. Já a concordância esperada (p_e) é dada por:

$$p_e = \left[\left(\frac{n_1}{n}\right) \times \left(\frac{m_1}{n}\right)\right] + \left[\left(\frac{n_0}{n}\right) \times \left(\frac{m_0}{n}\right)\right]$$

Substituindo pelos valores da Tabela 3, tem-se [(20/100) * (25/100)] + [(75/100) * (80/100)]. Logo, p_e será igual a 0,65. Para chegar ao valor final de Kappa (*k*), o cálculo deverá seguir a equação:

$$k = \frac{(p_o - p_e)}{(1 - p_e)}$$

A estatística Kappa (k) do exemplo, portanto, é de 0,57. Apenas com esse índice, entretanto, não se pode tirar conclusões. O próximo passo é compreender o que este índice representa.

Landis e Koch (1977) definiram classes para o grau de concordância, a fim de manter uma nomenclatura consistente ao descrever a força da confiança associada ao Kappa (LANDIS; KOCH, 1977). Tais categorias também foram utilizadas por diversas pesquisas, inclusive a de Freitas e Vieira (2013) na anotação de um corpus em português contendo opiniões sobre filmes.

Landis e Koch (1977) sugerem que os resultados do Kappa de Cohen sejam interpretados conforme a Tabela 4. Na primeira coluna, observam-se os limites mínimos e máximos de cada nível correspondente à direita. A escala começa no nível

Fraco e atinge o Quase Perfeito, ideal em atividades de anotação (LANDIS; KOCH, 1977).

Tabela 4 – Classes de concordância do coeficiente Kappa

Índice kappa	Grau de Concordância
< 0.00	Fraco
0,00 – 0,20	Leve
0,21 – 0,40	Razoável
0,41 – 0,60	Moderado
0,61 – 0,80	Substancial
0,81 – 1,00	Quase Perfeito

Fonte: adaptado de Landis e Koch (1977)

Porém, McHugh (2012) questiona os limites desses intervalos. Para ele, estes níveis permitem que pouco acordo entre anotadores seja classificado como substancial. 61% de concordância já pode ser visto como problemático, já que quase 40% dos dados representam informações defeituosas. “Para laboratórios clínicos, ter 40% das avaliações de amostras erradas, seria um problema de qualidade extremamente sério” (MCHUGH, 2012).

Por isso, McHugh (2012) complementa dizendo que muitos textos recomendam o mínimo de 80% de concordância entre anotadores. Em seguida, a Tabela 5 mostra a proposta de McHugh (2012), corrigindo os intervalos e categorias nominais. Em resumo, qualquer nível abaixo de 0,60 é considerado inadequado.

Tabela 5 – Classes de concordância do coeficiente Kappa corrigidas

Índice kappa	Grau de Concordância
0,00 – 0,20	Nenhum
0,21 – 0,39	Mínimo
0,40 – 0,59	Fraco
0,60 – 0,79	Moderado
0,80 – 0,90	Forte
> 0,90	Quase Perfeito

Fonte: adaptado de McHugh (2012)

“Em linguística computacional, o limite de aceitabilidade do grau de concordância de um corpus anotado pode variar de pesquisador para pesquisador” (DOSCIATTI; FERREIRA; PARAISO, 2015, p.125). Para Krippendorf (1980 apud Dosciatti, Ferreira e Paraiso, 2015), uma anotação só pode ser aceita em cenários

onde o Kappa é maior que 0,67. Seguindo as classes de Landis e Koch (1977) (vide seção 3.2.2), Freitas e Vieira (2013) consideraram como adequado o valor de Kappa igual ou superior a 0,39. Eugenio e Glass (2004) afirmam que, ao invés de se basear em métricas pré-definidas, é necessário avaliar a metodologia de anotação, seu nível de detalhamento e quais diretrizes foram seguidas.

Como afirma Fleiss (1971), houve variações e tentativas de expandir os casos nos quais o Kappa poderia ser usado. Outra variação de Kappa de Cohen é conhecida como Kappa ponderado (do inglês, *Weighted* Kappa). Este modelo é útil quando se deseja atribuir pesos diferentes na discordância entre categorias (FLEISS, 1971). Fleiss (1971), na tentativa de generalizar o modelo do Kappa de Cohen, modificou as fórmulas originais com o propósito de englobar casos onde há três ou mais anotadores.

Há outros métodos para avaliar a concordância entre observadores, embora o coeficiente Kappa seja a medida mais comum relatada na literatura para esse fim, principalmente a médica (VIERA; GARRETT, 2005). Recentemente, alguns estudos também o utilizaram na anotação de corpus em português (DOSCIATTI; FERREIRA; PARAISO, 2015); (FREITAS; VIEIRA, 2013).

Com o embasamento teórico exposto até aqui, propõe-se um modelo de anotação de textos de opiniões em português. Para isso, é desenvolvida uma ferramenta para facilitar a obtenção dos exemplos etiquetados. O próximo capítulo a apresentará, introduzindo também, o corpus de *reviews* utilizado.

4 PROPOSTA DE UM MODELO PARA ANOTAÇÃO DE *REVIEWS*

Para o desenvolvimento da ferramenta de etiquetação de opiniões, faz-se necessário primeiramente, a escolha pela criação de um novo corpus ou utilização de corpora já construídos por outros autores. Em pesquisa bibliográfica inicial, foi constatado que há conjuntos de *reviews* sobre produtos extraídos de *websites* em português. A seção 4.1 apresenta as características principais do corpus criado por Hartmann et al. (2014), o qual foi selecionado para análise nesta pesquisa.

4.1 APRESENTAÇÃO DO CORPUS

Com a disseminação dos dispositivos computacionais e *mobiles*, qualquer um é capaz de postar comentários na Web (HARTMANN et al., 2014). Nesses casos, a linguagem formal não é necessariamente utilizada. Aliás, o que marcam esses textos são justamente as gírias da Internet, com repetição de vogais, expressões coloquiais, abreviações, entre outras estruturas sintáticas não complexas (SQUIRES, 2010). Segundo Hartmann et al. (2014), isso afeta severamente o campo do Processamento de Linguagem Natural (PLN), tanto para fins de análise linguística, quanto ML, que requerem textos bem escritos. O que é o PLN e como suas técnicas podem ser aplicadas são abordadas, com mais detalhes, no capítulo 6.

Tendo isso em vista, Hartmann et al. (2014) propuseram uma ferramenta capaz de normalizar lexicamente opiniões retiradas da Internet, ou seja, corrigir palavras mal formadas, fora do vocabulário comum. Para tal, os autores construíram um corpus de opiniões sobre produtos do Buscapé[1], um site no qual é possível postar vantagens e desvantagens sobre diversos produtos, serviços e empresas (HARTMANN et al., 2014).

Ao total, foram extraídas 85.910 *reviews* sobre 16.667 produtos diferentes da base de dados do Buscapé. Ele foi escolhido para análise de spams de opinião pelos seguintes motivos: 1) ser completamente em português; 2) ter um grande número de *reviews*; 3) acesso na íntegra aos textos das opiniões e metadados sobre as avaliações, como código do usuário, data da postagem e *rating*. O *rating* das *reviews*

[1] Site do Buscapé disponível em: www.buscape.com.br

se dá pelo número de estrelas (de uma a cinco), especificado pelo usuário que a postou. O Quadro 3 lista todos os atributos presentes sobre as opiniões.

Quadro 3 – Atributos sobre o corpus do Buscapé

Atributo	Descrição
Source	URL de onde a review foi extraída
Evaluation_date	Dia em que a review foi postada
Category	Categoria do produto avaliado
Product	Nome completo do produto
Stars	Número de estrelas dadas pelo usuário que avaliou o produto
Recommends	Indicador que aponta se o usuário recomenda a compra do produto ou não
User	Código numérico que identifica o usuário que avaliou o produto
ThumbsUp	Quantidade de usuários que gostaram da avaliação
ThumbsDown	Quantidade de usuários que não gostaram da avaliação
Pros	Pontos positivos apontados pelo usuário sobre o produto
Cons	Pontos negativos apontados pelo usuário sobre o produto
Opinion	Texto na íntegra da opinião escrita pelo usuário

Fonte: elaborado pelo autor

A Figura 8 apresenta um extrato do corpus de Hartmann et al. (2014) em XML, no qual é possível visualizar todos os atributos dispostos no Quadro 3. O endereço completo do *website* onde a *review* se encontra foi ocultado para não expor o nome e a conta do usuário.

Figura 8 – Extrato do corpus do Buscapé

```
<review>
  <source value="http://www.buscape.com.br/
  24--56256.html"/>
  <evaluation_date value="06/10/2009"/>
  <category value="Bicicleta"/>
  <product value="Bicicleta Track                "/>
  <stars value="5.0"/>
  <recommends value="Yes"/>
  <user value="3489"/>
  <thumbsUp value="1"/>
  <thumbsDown value="1"/>
  <pros>Desempenho Qualidade Recursos Adicionais </pros>
  <cons>Não possui nenhum ponto negativo. </cons>
 −<opinion>
    "Quando decidi que era este o produto, eu já estava satisfeita com as suas funcionalidades e passei a comparar
    preço nas lojas. Quando recebi em casa, o produto me encantou ainda mais." O que gostei: Bonita, leve, designer
    moderno e vem com alguns acessórios. O que não gostei: Produto não vem com uma bomba para encher os
    pneus.
  </opinion>
</review>
```

Fonte: adaptado de Hartmann et al. (2014)

Além do texto da opinião propriamente dito, essencial para análises linguísticas, há também metadados sobre a *review*, os quais podem ser úteis para estudos comportamentais. Assim, tem-se recursos para abordar as duas vertentes descritas na seção 2.2 por Sandulescu e Ester (2015).

Santos, Júnior e Camargo (2018) apontam a necessidade de reduzir a quantidade de dados a serem classificados dependendo do número limitado de anotadores. Com isso, algumas medidas tiveram que ser tomadas.

Em média, cada produto possui 5,15 *reviews* associadas a ele, apesar de 65% das mercadorias possuírem entre 1 e 4 avaliações. Isso pode ser explicado pelo fato de que há produtos consideravelmente populares, os quais chegam ao máximo de 230 opiniões relacionadas. Com o intuito de não criar um subconjunto de *reviews* tendenciosas, estabeleceu-se algumas premissas: se encontrados possíveis spams em produtos com poucas *reviews*, o impacto será alto. Se encontrados spams em produtos com muitas *reviews*, o impacto será mínimo. Logo, buscou-se selecionar todas as opiniões sobre mercadorias balanceadas, as quais obtivessem um total de 5 *reviews* cada. Equivalente à média de avaliações por produto.

Do subconjunto obtido (16.375), foram excluídas avaliações cujos textos fossem compostos de nenhum ou apenas um caractere, por não representar palavras ou frases. Além disso, algumas opiniões foram excluídas por conterem somente uma ou duas palavras que não apresentassem sentido ou pela repetição de letras consecutivas. Os produtos que ainda possuíam 5 *reviews* vinculadas, mesmo após a aplicação desses filtros, totalizaram 3.028, ou seja, 15.140 opiniões (17,6% do corpus).

Buscando complementar os trabalhos existentes, principalmente o de Costa, Benevenuto e Merschmann (2013), que estudaram *reviews* na língua portuguesa, verificou-se a necessidade de classificar *reviews* em termos de experiência do usuário com o produto. Significa identificar se o postador tem ou não experiência com a mercadoria avaliada por ter expressado que possui ou não o produto, ter utilizado ou, até mesmo conhecido alguém que o tenha feito. Apesar desses tipos de opinião não necessariamente representarem que os indivíduos deliberadamente tentaram favorecer ou prejudicar seus alvos, Liu (2012) argumenta que, por não possuírem experiência prévia com o produto, sua opinião não é genuína.

Esta categoria de opinião, embora represente apenas uma fração do problema identificado por Jindal e Liu (2008), ainda é um obstáculo que pode estar afetando consumidores na hora de adquirir produtos. Já que spams ainda não foram analisados por esta perspectiva.

Com base nas *reviews* do Buscapé, a seção seguinte introduz a Anoto-PT[2], modelo de ferramenta desenvolvido neste trabalho para a obtenção de anotações manuais.

4.2 ANOTO-PT

Mukherjee e Venkataraman (2014) elencam que obter anotações confiáveis em larga escala não é trivial, além de ser uma tarefa custosa e que consome tempo. Drury et al. (2014) argumentam que a anotação manual de dados é uma tarefa repetitiva e que a contratação de anotadores em tempo integral para realizar essa atividade pode ultrapassar o orçamento disponível. Santos, Júnior e Camargo (2018), por exemplo, convocaram dois voluntários para anotação de 3.000 textos para realização de tarefas relacionadas à Análise de Sentimentos. Eles finalizaram essa atividade após 3 meses.

Em contrapartida, a abordagem *crowd-sourcing* tem sido amplamente utilizada para evitar que pessoas sejam contratadas para realizar tais atividades (DRURY et al., 2014). Nela, tarefas de anotação são distribuídas em um grande número de indivíduos que passarão menor tempo anotando dados (DRURY et al., 2014). Wang, Hoang e Kan (2013, p. 10, tradução nossa) elucidam que "*crowd-sourcing* é uma estratégia que combina o esforço do público para resolver um problema ou produzir um recurso". Entretanto, a qualidade das etiquetações é inversamente proporcional à quantidade de anotadores. Isso porque pode haver uma queda no entendimento das regras de etiquetação, além de atrair pessoas que não são especialistas sobre as informações sendo classificadas (WANG; HOANG; KAN, 2013).

Buscou-se combinar as abordagens *crowd-sourcing* e utilização de especialistas para a proposição da Anoto-PT, na tentativa de unir os pontos positivos de ambas as metodologias. Algo, até então, não identificado em trabalhos anteriores. O *crowd-sourcing* não foi implementado em sua totalidade, visando ter maior

[2] Site da Anoto-PT disponível em: www.tcfernando.com.br ou via código-fonte em: www.github.com/FernandoSchuch/anoto-pt

qualidade nas anotações, embora isso possa ter impacto no número de opiniões classificadas. Já os voluntários foram convocados devido à proximidade com o autor dessa pesquisa e por já terem comprado produtos online, indicando certo nível de familiaridade com ambientes semelhantes ao Buscapé. A fim de torná-los especialistas no assunto, métodos de treinamento foram aplicados.

Segundo Wang, Hoang e Kan (2013), a motivação em participar de tarefas de anotação pode ser financeira ou altruística. O presente trabalho não ofereceu nenhum tipo de recompensa financeira aos participantes da etiquetação do corpus. Os voluntários convidados a colaborar com a pesquisa, portanto, decidiram participar somente pelo propósito altruístico.

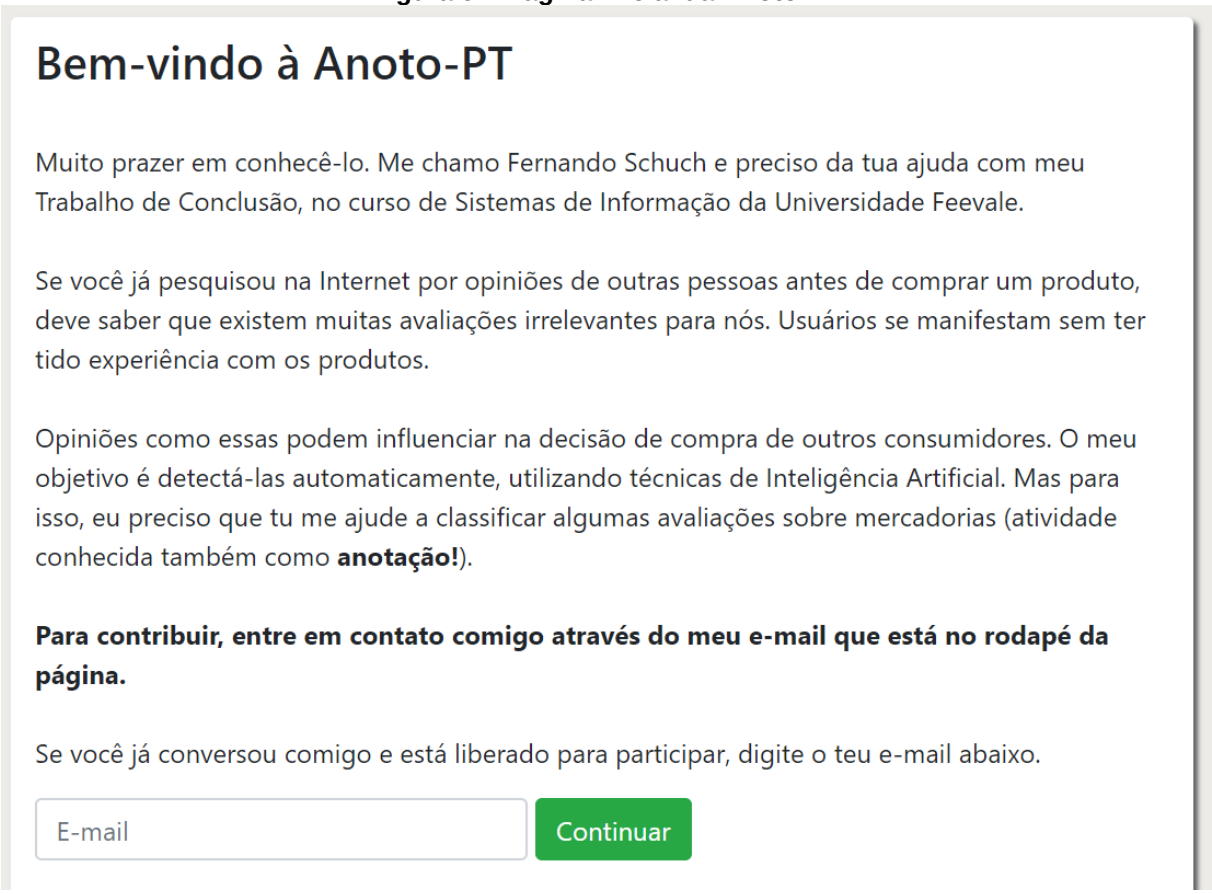
Levando em conta os aspectos descritos acima, foi proposta a criação de um *website,* objetivando aumentar a facilidade de acesso à Anoto-PT, já que o mesmo pode ser acessado de qualquer lugar, seja por computadores, notebooks ou *smartphones*, bastando ter uma conexão com a Internet.

O processo de anotação foi realizado em dois momentos. No primeiro, buscou-se realizar um projeto piloto da ferramenta (versão 1.0), com o propósito de entender o comportamento dos voluntários, suas dificuldades durante a atividade de etiquetação, além de obter resultados preliminares das anotações feitas por eles. A partir das conclusões obtidas, propôs-se o segundo momento. Nele, foram realizados ajustes pertinentes no projeto piloto, resultando em uma segunda versão da Anoto-PT. Com isso, uma nova bateria de anotações foi coletada. Ambas as versões são descritas nas próximas subseções, explicando a evolução da ferramenta.

4.2.1 Versão 1.0

A Figura 9 ilustra a página inicial da Anoto-PT. Ao acessar o site, há um texto introdutório, com o objetivo de apresentar o autor, a instituição de ensino onde a pesquisa foi desenvolvida, bem como contextualizar o problema dos spams de opinião. À título de simplificação, o termo “spams de opinião” não foi diretamente apresentado. Foi elucidado apenas, a necessidade em detectar avaliações que não representassem a real experiência de um usuário com o produto sendo avaliado por ele. Em seguida, é feito um pedido de colaboração para que o participante ajude através da anotação de algumas avaliações.

Figura 9 – Página inicial da Anoto-PT



Fonte: elaborado pelo autor

Para colaborar, o e-mail do voluntário é requisitado antes de passar para a próxima etapa. Isso foi necessário para evitar que a mesma pessoa anotasse mais de uma vez a mesma *review*. Além disso, caso o mesmo participante estivesse disposto a retornar ao site posteriormente para avaliar mais opiniões, bastava informar o mesmo e-mail para retomar o progresso. Não houve, neste primeiro momento, limite na quantidade de etiquetações que uma mesma pessoa poderia dar.

Ao informar o e-mail, instruções de como classificar cada opinião foram mostradas ao participante. Para cada *review*, o voluntário deveria ler o texto representando a opinião e o nome produto ao qual ela se refere. Em seguida, escolher uma das cinco opções de classificação que melhor descreve a experiência do usuário com o produto sobre o qual está opinando. São elas:

a) Possui ou conhece alguém que possui o produto;

b) Não possui o produto, mas deseja comprá-lo;

c) Não possui o produto;

d) Não há como afirmar; e

e) A opinião não está relacionada ao produto.

As três primeiras possibilidades de classificação (*a*, *b* e *c*) remetem diretamente ao problema de possuir ou não o produto. Além de ter a mercadoria, o usuário poderia conhecer um terceiro que o possui. Portanto, este cenário também foi considerado como sendo uma opinião genuína. A possibilidade de não possuir o produto foi dividida em duas opções (*b* e *c*). O que difere uma da outra, é o desejo do usuário que postou a *review* comprar o item posteriormente. Pensou-se que essa discriminação poderia auxiliar os voluntários a escolher a opção mais adequada de classificação.

Após leitura de parte das *reviews* do corpus do Buscapé, foi identificado que há textos em que não é possível afirmar se o indivíduo tem ou não o produto. A Figura 10 demonstra um exemplo de opinião que se caracteriza nesta opção.

Figura 10 – Exemplo de opinião "não é possível afirmar"

> Fui muito bem atendido. Valeu pelo esforço de vocês.
>
> O que gostei: estou muito feliz, mas comprar assim é tanto que complicado pois os tamanhos nem sempre batem com o que se espera.
>
> O que não gostei: não tenho certeza, mas acho que o número 10 nas costas está meio torto.

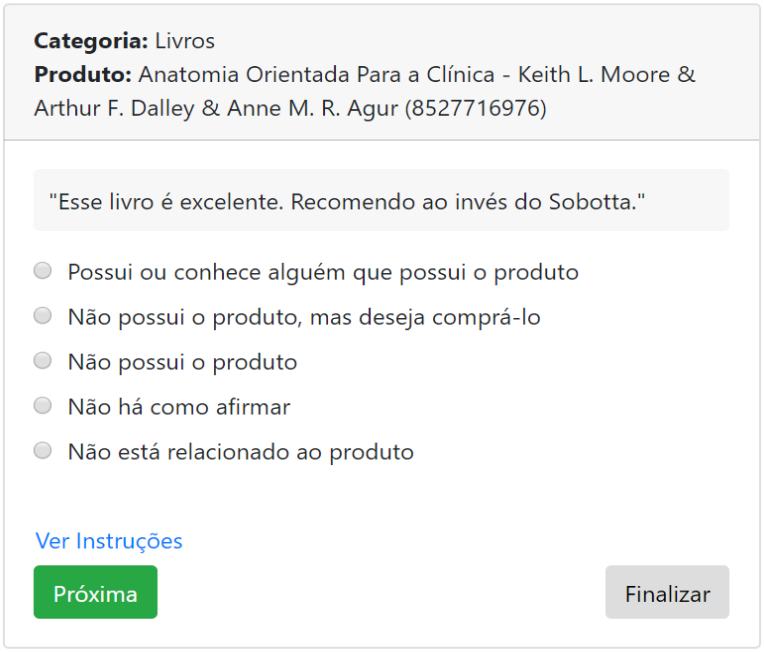
Fonte: elaborado pelo autor

No texto acima, não há como afirmar se quem postou a *review* adquiriu a mercadoria. Por esse motivo, a quarta opção (d) também está contida na lista de possibilidades. Se o anotador tivesse qualquer dúvida sobre qual das três primeiras opções marcar, poderia escolher esta resposta.

Costa, Benevenuto e Merschmann (2013), em sua pesquisa de opiniões no Apontador, elucidaram que há *reviews* que possuem conteúdo irrelevante, ou que não estão relacionados ao alvo sendo avaliado. Por isso, foi considerado importante adicionar uma quinta possibilidade de resposta (*e*), mesmo que ela não seja aprofundada neste trabalho. Ela abrange spams dos tipos 2 e 3 de Jindal e Liu (2008) (vide seção 2.1), onde não há relação entre o texto da *review* com as características do produto sendo avaliado. Por exemplo, usuários falando de um produto A, mas

postando na página do produto B. Ou ainda, transcrevendo um exemplo de *review* extraída do corpus do Buscapé: “Fui roubada, eu e meus dois filhos. Levaram nossos Iphones”. Na Figura 11, é possível visualizar a página de anotação.

Figura 11 – Página de anotação da Anoto-PT



Fonte: elaborado pelo autor

Wang, Hoang, Kan (2013) explicam que anotações feitas por especialistas nos dados sendo etiquetados resultam em anotações de melhor qualidade. Tendo isso em vista, os voluntários selecionados seguiram o pré-requisito de já terem pesquisado por opiniões online antes de tomar decisões de compra. Embora isso não os caracterize como especialistas, os participantes escolhidos seguem a premissa de já ter tido contato com sites de opiniões. Os voluntários foram incentivados a convidar outras pessoas que também estivessem dispostas a participar da pesquisa, seguindo os pré-requisitos iniciais.

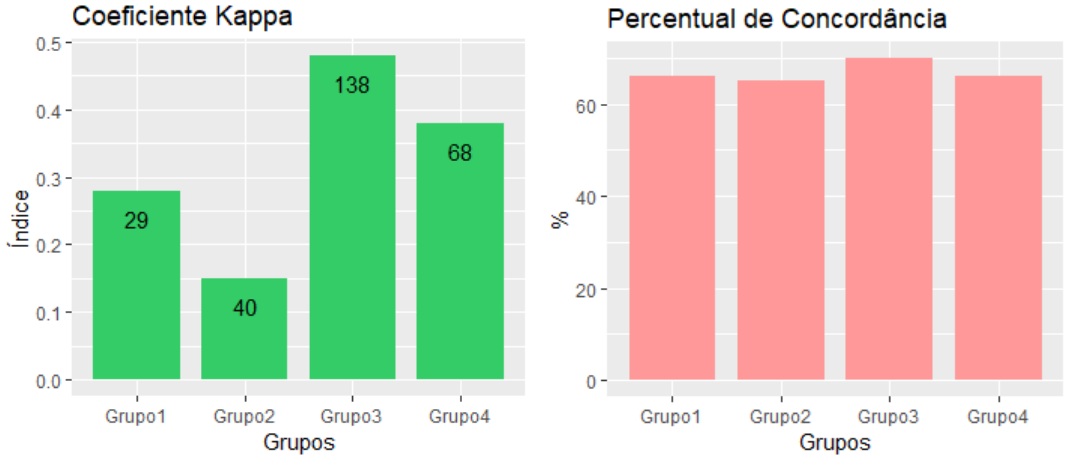
Neste projeto piloto, os participantes poderiam anotar quantas *reviews* desejassem. Com isso, foi possível identificar o nível de comprometimento dos voluntários. Cada pessoa anotou, em média, 38 *reviews* que apareciam de modo aleatório, totalizando 1.099 etiquetações entre os dias 9 e 18 de Abril de 2019.

Visando aplicar o coeficiente Kappa de Cohen, foram selecionadas as duplas de participantes que mais anotaram *reviews* em comum, a fim de avaliar o nível de concordância entre eles. Foram selecionados quatro grupos. O grupo com menos anotações de *reviews* em comum foi o Grupo 1, com 29 classificações, enquanto o

maior foi o Grupo 3 com 138 etiquetações. Os resultados obtidos de concordância podem ser visualizados no gráfico da Figura 12.

Enquanto o menor percentual simples de confiança calculado foi de 65% para o Grupo 2, o Kappa de Cohen atingiu o índice de apenas 0,13. O máximo obtido entre os dois métodos é observável no Grupo 3. O percentual de concordância atingiu 70%. Já o Kappa, 0,48.

Figura 12 – Resultados do coeficiente Kappa e percentual de concordância



Fonte: elaborado pelo autor

É possível concluir que o coeficiente Kappa apresenta uma confiança muito menor do que a calculada pelo percentual de concordância simples. Seguindo as métricas expostas por Landis e Koch (1977) introduzidas na seção 3.2.2, apenas o Grupo 3 atingiu o grau Moderado, enquanto os Grupos 1 e 4, chegaram apenas a Razoável. Já o segundo grupo, foi classificado como Leve. Entretanto, se utilizarmos os limites de McHugh (2012), nenhum deles ultrapassou o nível Fraco.

Os resultados iniciais indicam a dificuldade de duas pessoas em chegar a um consenso sobre o que determinadas opiniões expressam. O baixo grau de concordância obtido entre os anotadores pode ter sido causado por dois fatores. Primeiramente, não houve treinamento prévio para os participantes. Isso implica na possibilidade de que eles possam não ter entendido alguma instrução antes de iniciar a tarefa ou precisaram de algumas anotações para se familiarizarem com a ferramenta.

O segundo fator é justamente a similaridade entre as opções para anotação “Possui o produto” e “Não há como afirmar”. Mais de 43% das discordâncias

ocorreram por um anotador marcar que o usuário possuía um produto, enquanto o segundo assinalava que não era possível afirmar.

A próxima subseção descreverá as melhorias propostas na segunda versão da ferramenta.

4.2.2 Versão 2.0

A partir das conclusões observadas na versão piloto, fez-se necessário implementar ajustes na Anoto-PT para aumentar a qualidade das anotações. A nova versão foi disponibilizada para etiquetação manual de 26 de Agosto até 28 de Setembro de 2019. Nela, foi adicionado um vídeo tutorial de 3 minutos de duração[3], apresentado aos voluntários antes de iniciar a anotação da primeira *review*.

O tutorial foi proposto com o intuito de substituir um treinamento presencial, visto que poderia tomar tempo dos participantes. Além disso, alguns indivíduos podem se encontrar distantes geograficamente. No vídeo, explicações mais detalhadas sobre as opções de etiquetação, bem como a utilização da ferramenta eram explicados ao anotador. “Essas instruções buscam nivelar o conhecimento dos anotadores, em uma tentativa de reduzir a subjetividade em tarefas de anotação” (SANTOS; JÚNIOR; CAMARGO, 2018, p. 297, tradução nossa).

Em alguns cenários, o usuário que postou a *review* não deixava claro se ele tinha ou não o produto. Entretanto, era perceptível que ele possuía experiência com o mesmo. A Figura 13 apresenta dois exemplos de opiniões com característica distintas, as quais deixaram os anotadores em dúvida sobre suas classificações.

Figura 13 – Usuário que possui o produto x Usuário com experiência

Opinião 1	*Opinião 2*
Tenho este aparelho de pressão e ele é um dos melhores do mercado pois mede com muita precisão.	No geral é um excelente produto, cumpre o que promete.

Fonte: elaborado pelo autor

Na primeira opinião, percebe-se que quem postou a *review* parece ter o produto, enquanto na segunda, não há como afirmar isso. Embora, pode-se dizer que

[3] Vídeo tutorial disponível em: www.youtube.com/watch?v=4KXjNDV37b4

ambos os usuários têm experiência com ele. Portanto, as opções de etiquetação citadas na seção 4.2.1 foram reformuladas:

a) Possui, utilizou ou conhece alguém que possui ou utilizou produto;

b) Não possui nem utilizou o produto;

c) Não há como afirmar;

d) Não está relacionado ao produto.

O intuito dessas novas classes foi tornar a etiquetação mais abrangente. O postador não necessariamente deveria ter a mercadoria, mas sim apresentar sinais de ter experiência com o produto, seja por ter comprado, utilizado ou pelo menos ter descrito a experiência de outra pessoa. As duas opções relacionadas a não possuir o produto foram unificadas (vide seção 4.2.1). Não foram encontradas fundamentações nem propósitos para mantê-las discriminadas.

A dinâmica de anotação também foi repensada para maximizar o número de *reviews* anotadas por produto, objetivando a análise de impacto no *rating* dos mesmos. Os primeiros produtos escolhidos aleatoriamente para anotação deveriam ter suas 5 *reviews* etiquetadas antes de buscar as próximas mercadorias ainda não anotadas. Isso permite que seja possível aprofundar o estudo de número de *reviews* nos produtos, ou até mesmo identificar quais categorias de mercadorias foram mais afetadas.

Os anotadores também não poderiam etiquetar quantas *reviews* quisessem. Na versão 2, foram liberadas quantidades limitadas de opiniões para os voluntários, com o intuito de realizar o cálculo de concordância na totalidade das anotações, algo que não aconteceu na primeira versão pela disparidade no número de anotações entre os participantes.

O Kappa de Cohen especifica que seu cálculo deve ser feito sobre o mesmo número de anotações realizadas entre dois jurados distintos. Portanto, as duplas foram sendo formadas à medida que novos voluntários eram convidados a participar. Cada dupla poderia avaliar um número diferente de *reviews*. Os resultados das anotações são apresentados e discutidos no próximo capítulo.

5 RESULTADOS DA ANOTAÇÃO

A Tabela 6 descreve algumas estatísticas sobre as versões da ferramenta. É possível identificar que a participação dos voluntários durante a versão piloto foi maior. Isso porque, em média, cada participante anotou 3,8 opiniões por dia, enquanto na segunda versão, esse número reduziu para 2 anotações/dia.

Esse comportamento pode ser explicado pelo limitador na quantidade de *reviews* a serem etiquetadas. Na primeira versão, não havia limite de contribuições possíveis e nem indicativo de quantas ele já havia anotado, por isso, o participante continuava etiquetando quantas opiniões ele se sentisse confortável. Já na 2.0, além de mostrar quantos textos foram lidos e etiquetados, havia um número limitado de *reviews* disponíveis. Uma vez que as finalizassem, não houve proatividade por parte dos participantes em requisitar mais opiniões para anotação. Cabia ao autor da pesquisa, contudo, incentivá-los a continuar contribuindo.

Tabela 6 – Comparações entre as versões da Anoto-PT

Versão	Anotações	Voluntários	Período (dias)
1.0	1.099	29	10
2.0	1.560	23	34

Fonte: elaborado pelo autor

Foi percebido que o esforço empregado pelo autor para incentivar os voluntários com etiquetações pendentes a acessar a ferramenta foi maior durante a segunda versão. A queda no nível de contribuições também pode ser explicada pelo fato de que 11 dos 29 voluntários que participaram na primeira versão, também o fizeram na segunda. Ao contatá-los via ligação telefônica, e-mail ou por aplicativos de troca de mensagens instantâneas, percebeu-se que havia maior resistência desses voluntários em voltar a acessar a Anoto-PT. Logo, evitou-se pedir a ajuda em demasia para estas pessoas, já que elas poderiam ficar incomodadas. Portanto, a versão piloto interferiu negativamente no número de etiquetações feitas na Anoto-PT 2.0.

Apesar da menor média de anotações diárias, ao final dos 34 dias em que a Anoto-PT 2.0 ficou disponível para acesso dos voluntários, foi possível obter 1.560 anotações realizadas por 23 pessoas diferentes.

Ressalta-se que 44,8% das pessoas que contribuíram durante a primeira versão, foram convidadas pelos próprios voluntários. Nesses casos, não houve

interferência do autor da pesquisa. Já na Anoto-PT 2.0, essa estatística caiu para 30,4%. Isso demonstra que após utilizarem a ferramenta, os participantes sentiram-se dispostos a disseminar o *link* para que outras pessoas também colaborassem.

As etiquetações adquiridas na primeira versão da ferramenta não foram aproveitadas para análise de resultados, dado que houve alterações significativas na formulação das questões para a segunda versão, principalmente na quantidade e significados das opções de rotulação. Além disso, o treinamento por vídeo foi aplicado somente na Anoto-PT 2.0. Portanto, por não terem seguido os mesmos critérios, as anotações da 1.0 foram desprezadas.

Com base nas opiniões etiquetadas na Anoto-PT 2.0, as duas próximas seções analisam os níveis de concordância entre os anotadores, bem como o impacto dos spams no *rating* dos produtos.

5.1 GRAU DE CONCORDÂNCIA DOS ANOTADORES

O capítulo 3 elucidou a importância em dar confiabilidade à anotação manual de corpora, utilizando métricas que aferem o grau de concordância entre jurados. Esta seção apresenta o cálculo do coeficiente Kappa de Cohen, a fim de selecionar apenas as *reviews* anotadas que ultrapassaram um valor mínimo estabelecido como aceitável.

Assim como na primeira versão, as pessoas que atenderam aos critérios de seleção explicados na seção 4.2, passaram a ser convocadas a colaborar. À medida que os possíveis voluntários aceitavam participar da pesquisa, eles foram sendo agrupados com o indivíduo imediatamente anterior ou próximo que também concordou em contribuir para o trabalho. Não houve critério referente à ordem de convocação dos voluntários para formação das duplas, dado que reunir pessoas com características aleatórias pode representar melhor a realidade de possíveis consumidores, os quais possuem percepções distintas.

Além disso, seria necessário investigar mais a fundo as experiências prévias das pessoas com sites de venda online, para que o agrupamento fosse adequado. Por esse não ser o objetivo do trabalho, preferiu-se não estabelecer uma ordem, embora a inserção desse estudo poderia aumentar a qualidade das anotações.

Durante a coleta de etiquetações, foram formadas 14 duplas de voluntários, denominados como grupos. Uma mesma pessoa poderia estar contida em vários

grupos, já que alguns participantes se dispuseram ou foram convidados a anotar mais *reviews* em diferentes momentos.

Através das publicações elencadas na seção 3.2.2, onde buscou-se verificar o que trabalhos relacionados consideram adequado para o grau de concordância, foi possível averiguar que não há definição clara sobre valores toleráveis em anotações de corpora. Um grau alto implica na redução do número final de *reviews* válidas. Já se o valor mínimo for muito baixo, etiquetações de má qualidade podem predominar.

Para esta pesquisa, foi estabelecido que o grau mínimo de concordância que deve ser atingido pelas duplas de anotadores é de 0,41. De acordo com as classes definidas na Tabela 4 por Landis e Koch (1977), isso indica que o nível de acordo entre os voluntários é tido como moderado. As demais referências aos níveis do Kappa também seguiram os intervalos dispostos na Tabela 4.

Na Tabela 7, verificam-se os resultados referentes às etiquetações de cada grupo, através do total de *reviews* anotadas, bem como em quantas eles concordaram. A partir das estatísticas dispostas, foi possível identificar o índice de Kappa (*k*). O cálculo do coeficiente foi implementado na Anoto-PT, com o objetivo de reduzir o tempo gasto na tarefa de criar a matriz de confusão, visto que é necessário dispor, uma a uma, todas as respostas na matriz (vide seção 3.2.2).

Tabela 7 – Resultados das anotações

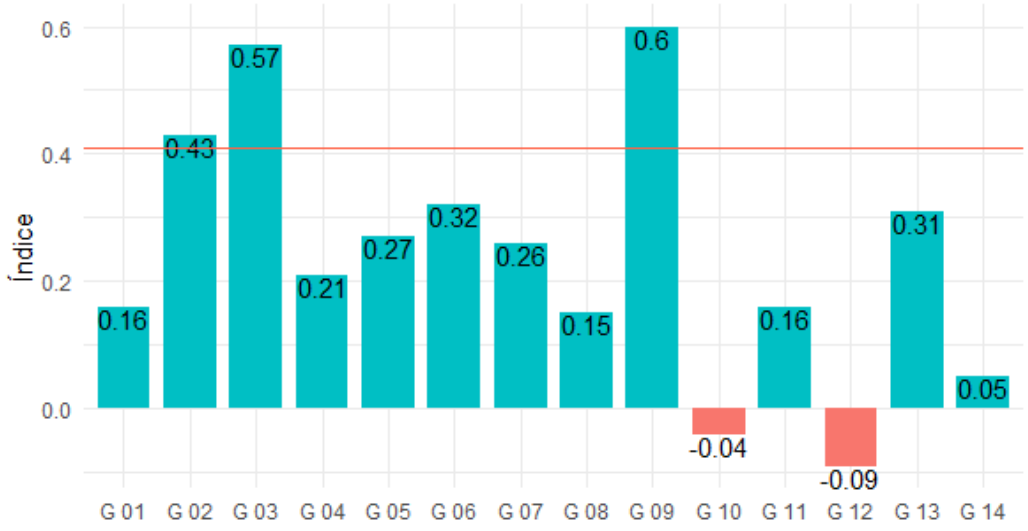
Grupo	Total *reviews* anotadas	Percentual Simples	Kappa
1	90	65,6%	0,16
2	130	70%	0,43
3	60	83,3%	0,57
4	50	62%	0,21
5	50	52%	0,27
6	50	80%	0,32
7	50	64%	0,26
8	50	52%	0,15
9	80	82,5%	0,60
10	50	14%	-0,04
11	30	36,7%	0,16
12	30	20%	-0,09
13	30	83,3%	0,31
14	30	30%	0,05
Total:	**780**		

Fonte: elaborado pelo autor

Assim como foi identificado nos resultados da Anoto-PT 1.0, um alto percentual de concordância simples não significa que o *k* será elevado. Os dois integrantes do Grupo 1, por mais que tenham concordado em 65,6% das etiquetações, obtiveram Kappa igual a 0,16 (Leve). No Grupo 13, o mesmo cenário é identificado. Contudo, conseguiram chegar em 0,31 (Razoável).

A Figura 14 apresenta um gráfico com os resultados de *k* para cada grupo. A linha horizontal, em laranja, representa o valor mínimo que deveria ser atingido para ser considerado como uma etiquetação confiável. Percebe-se que apenas 3 das 14 duplas ultrapassaram o índice aceitável. Destaque para o Grupo 9, que teve seu *k* igual a 0,60, enquanto a pior das duplas chegou a 0,09 negativos (Grupo 12).

Figura 14 – Valores de *k* em cada grupo



Fonte: elaborado pelo autor

Os resultados visualizados na Figura 14 elucidam a dificuldade em conseguir anotações de qualidade. Embora as *reviews* rotuladas pelos grupos que não ultrapassaram o Kappa mínimo de 0,41 tenham que ser descartadas, foi possível identificar as duplas de voluntários que mais concordaram entre si. Próximas rodadas de anotação, por exemplo, poderiam alocar opiniões somente para os grupos 2, 3 e 9, ou ainda, investir em treinamentos mais intensos para os demais participantes.

Após desconsiderar os grupos nos quais os anotadores não chegaram ao nível aceitável de concordância, restaram 270 *reviews* válidas para o prosseguimento deste estudo. As demais opiniões foram descartadas. Isso significa que apenas 34,6% das 780 opiniões podem ser utilizadas para posteriores modelagens computacionais.

Os índices calculados de Kappa mostram que o grau de concordância não foi perfeito em nenhum dos três grupos considerados, ou seja, houve indefinição sobre o que o texto de algumas *reviews* indicava, dentro das possíveis opções de etiquetação.

Seguindo a metodologia aplicada em Santos, Júnior e Camargo (2018), as opiniões cujos anotadores não chegaram à um consenso, passaram pela análise de um terceiro voluntário, que deu o seu voto de desempate. Ao todo, 53 opiniões tiveram que passar pela etiquetação de um terceiro jurado.

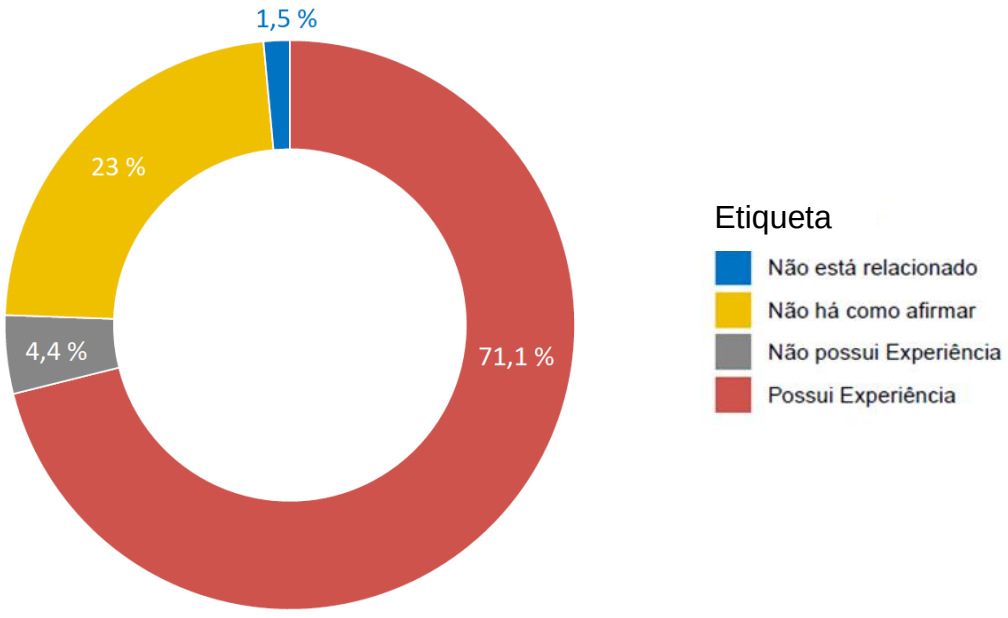
As 270 *reviews* estão dispersas entre 147 produtos de 53 categorias diferentes. Há, portanto, grande variedade de tipos de mercadorias presentes no corpus a ser estudado. Desde eletrodomésticos, como micro-ondas e geladeiras, passando pelos eletroeletrônicos, como câmeras digitais, notebooks e *smartphones,* até brinquedos e livros. A Tabela 8 e a Figura 15 elucidam, em números absolutos e percentuais, respectivamente, a distribuição das opiniões entre as quatro classes de etiquetação.

Tabela 8 – Distribuição absoluta das *reviews* por classe de etiquetação

Classe	Quantidade
Possui experiência	192
Não há como afirmar	62
Não possui experiência	12
Não está relacionado ao produto	4
Total:	270

Fonte: elaborado pelo autor

Figura 15 – Distribuição percentual das *reviews* por etiqueta



Fonte: elaborado pelo autor

A imensa maioria das opiniões classificadas afirmam que em 71,1% das *reviews*, o usuário possuía experiência com o produto sendo avaliado. Isso indica que, no mundo real, 28,9% das opiniões podem ser spams, apesar de que não seja possível garantir isso em determinados cenários. Como os 23% das *reviews* cujas etiquetas são “não é possível afirmar”, de acordo com os voluntários. Embora haja possibilidade de serem spams, houve insegurança por parte dos grupos de anotadores em escolher outras etiquetas.

O objeto de estudo desta pesquisa, entretanto, é observado nos 4,4% das *reviews* etiquetadas como “não possui experiência”. Nelas, os anotadores consideraram que o usuário não tinha o produto, nem mesmo o utilizou ou conhece um terceiro que tenha experiência. Logo, caracteriza-se dentro do tipo 1 de spams definidos por Jindal e Liu (2008).

Se este percentual fosse aplicado somente no corpus do Buscapé, haveria aproximadamente, 3.780 spams de opinião, os quais poderiam modificar o *rating* dos produtos. Claro que, este número ainda não considera os 1,5% das opiniões rotuladas como “não está relacionado ao produto”. Embora não seja abordado nesta pesquisa, este tipo de spam também contribui para prejudicar tomadas de decisão de compra (COSTA; BENEVENUTO; MERSCHMANN, 2013).

A seguir, na seção 5.2, será verificado como as *reviews* identificadas como spam afetam o *rating* final dos produtos.

5.2 IMPACTOS DOS SPAMS NO *RATING* DOS PRODUTOS

Selecionando apenas os 4,4% das *reviews* identificadas como spam de opinião, verificou-se a importância de analisar como elas podem modificar o *rating* dos produtos, expresso pelo número de estrelas. Das 147 mercadorias que tiveram ao menos uma opinião anotada pelos grupos concordantes, 10 mercadorias foram afetadas por avaliações de usuários que não tinham experiência com o mesmo. A Tabela 9 demonstra a variação dos *ratings* dos produtos após a remoção das *reviews* consideradas spams.

Importante ressaltar que, devido ao pré-processamento aplicado no corpus, explicado na seção 4.1, todos os produtos possuem 5 opiniões associadas. Com exceção da mercadoria Boneca, no qual foi identificado 3 opiniões spam, as demais

tiveram apenas uma *reviews* removida da composição do número médio de estrelas. Para as avaliações dos produtos elencados acima que não foram etiquetadas pelos anotadores, considerou-se que não eram spams. Os nomes reais dos produtos foram omitidos.

Tabela 9 – Rating inicial e final dos produtos

Produto	*Rating* Inicial	*Rating* Final	Diferença
Aparelho de Pressão	4	4,25	0,25
Bicicleta	4,8	4,75	- 0,05
Boneca	4,2	5	**0,80**
Camisa de Time de Futebol	4,4	4,25	- 0,15
Frigobar	3	3,25	0,25
Palm Top	4	3,75	**- 0,25**
Máquina de Costura	4,6	4,75	0,15
Micro-ondas	3,2	3,75	0,55
MP4 Player	2,8	3	0,20
Sofá	4	3,75	**- 0,25**
		Diferença média:	**0,15**

Fonte: elaborado pelo autor

Em média, o *rating* dos produtos da Tabela 9 foi elevado em 0,15 estrelas. A Boneca foi a mercadoria que teve maior aumento: 0,80. Enquanto dois produtos, o Palm Top e o Sofá, sofreram a queda máxima de 0,25 estrelas no *rating* final (destacados em negrito).

Com apenas 12 *reviews* de usuários que não possuíam experiência com o produto, é natural concluir que não seja possível realizar análises mais profundas, nem mesmo a criação de modelos computacionais para detecção automática de spams de opiniões. Baseando-se nessa premissa, foi identificada uma nova oportunidade para investigação.

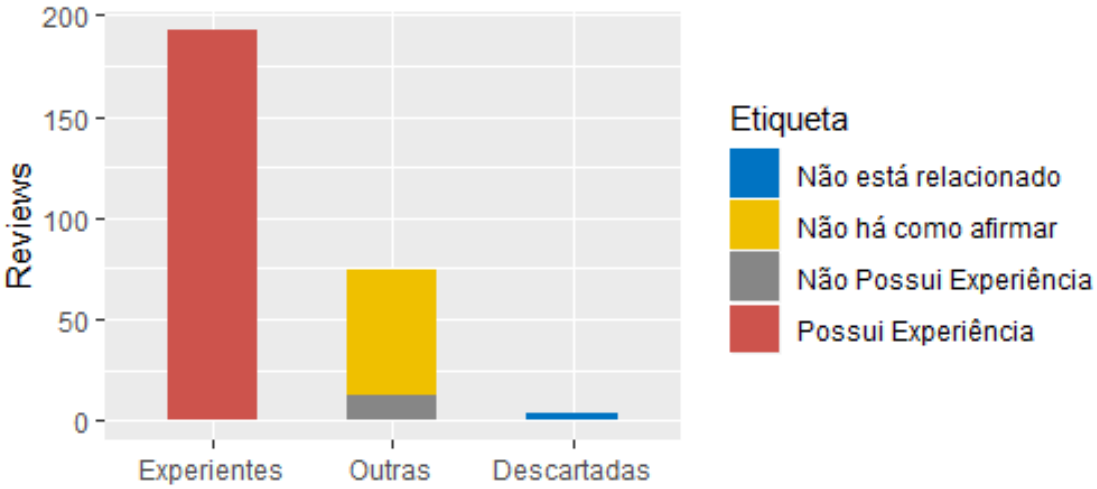
Ao invés de analisar textos que dão indícios da inexperiência de uma pessoa com determinada mercadoria, foram averiguadas opiniões nas quais é perceptível que o usuário realmente tem experiência sobre o produto no qual está opinando. De acordo com as etiquetações dos grupos concordantes, há 192 *reviews* anotadas neste cenário. Entendeu-se que a resolução do problema de poucos textos que serviriam como exemplo para treinamento de programas de ML, seria separar as opiniões de pessoas que têm experiência das que não têm ou não é possível afirmar.

Sendo assim, o intuito é possibilitar o balanceamento entre avaliações de usuários com experiência prévia com o produto e os demais tipos de avaliação. As

reviews em que o indivíduo não tem experiência com a mercadoria e aquelas em que não há como afirmar se o usuário possui ou pelo menos conhece o produto foram agrupadas. Embora não se enquadrem em nenhum tipo de spam definido por Jindal e Liu (2008), elas são ambíguas. Logo, há indícios de que o postador pode não ter fundamentos para argumentar sobre a qualidade do item sendo avaliado, apesar de não ser possível afirmar.

Portanto, as duas possibilidades de classificação de opiniões são nomeadas como: Experientes, ou seja, quem efetivamente teve contato prévio com o produto; e Outras, que englobam as demais opiniões, totalizando 192 e 74 *reviews* respectivamente. Lembrando que as opiniões etiquetadas como spam por não estarem relacionadas às características do produto, continuam de fora do estudo. O gráfico da Figura 16 ilustra o agrupamento descrito.

Figura 16 – Agrupamento de etiquetas por classe



Fonte: elaborado pelo autor

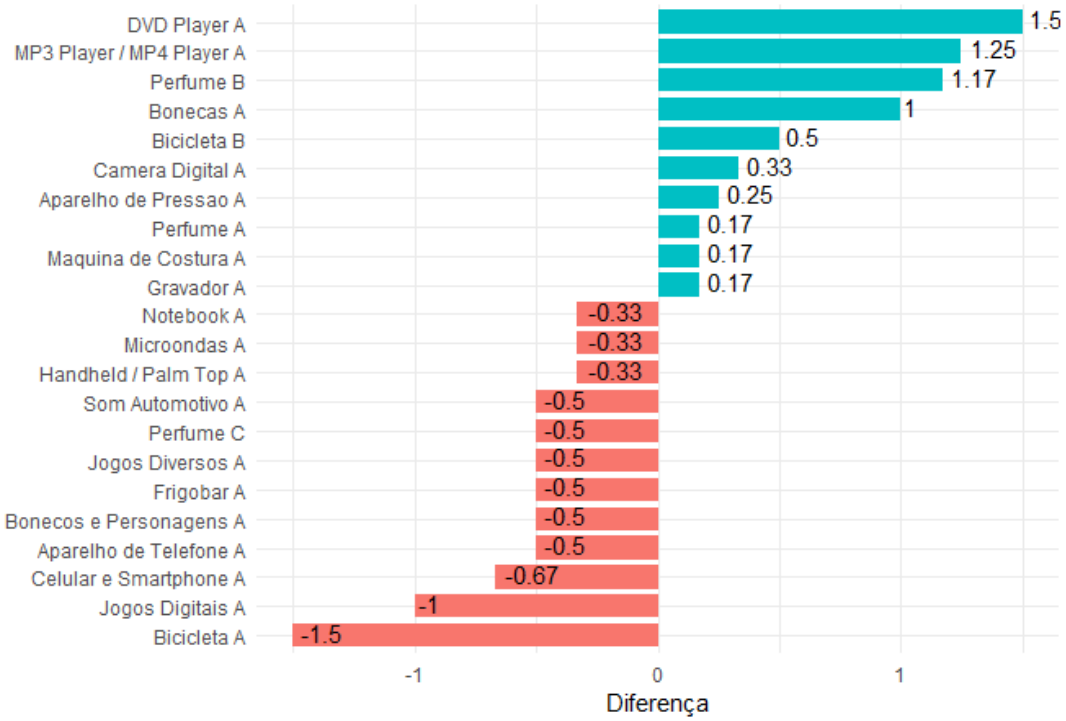
Após essas explicações, é importante reapresentar as novas estatísticas de impacto das opiniões da classe Outras no *rating* dos produtos. A Figura 17 resume a variação no número de estrelas. Para geração do gráfico, foram selecionadas apenas as *reviews* que foram efetivamente anotadas, ou seja, as avaliações não etiquetadas ou que foram anotadas por grupos não concordantes, não compuseram o cálculo do *rating* inicial das mercadorias. Após a remoção das opiniões do tipo Outras, 22 dos 147 produtos tiveram seu número de estrelas alterado.

Dos produtos afetados, 12 tiveram seu *rating* diminuído após a exclusão das *reviews* do tipo Outras. Para as mercadorias cujo *rating* aumentou, a média de crescimento foi de 0,65 estrelas. Variando de 0,17 até 1,5. Enquanto para os que

diminuíram, a média ficou em 0,60. A menor variação para esses casos foi de 0,33, enquanto a maior, 1,5.

Pode-se afirmar que há uma leve tendência em existir mais opiniões publicadas com o intuito de promover o produto, visto que 54,5% das *reviews* desconsideradas foram responsáveis por aumentar a reputação dos alvos.

Figura 17 – Variação no *rating* dos produtos



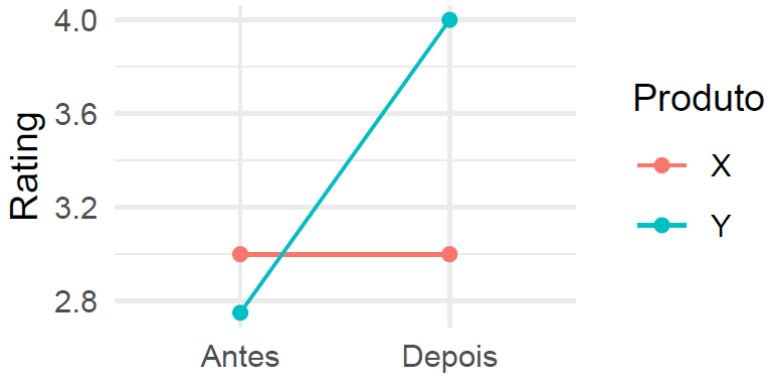
Fonte: elaborado pelo autor

A seguir, é apresentado um caso relevante sobre dois MP3 / MP4 *Players*, reprodutores comerciais de áudio e vídeo. Os nomes reais dos mesmos também foram omitidos.

Ao considerar todas *reviews* anotadas para as duas mercadorias, X e Y, elas tinham seus *ratings* iguais a 3,0 e 2,75, respectivamente. Entretanto, depois de examinar apenas as *reviews* do tipo Experiente, excluindo as Outras, a reputação do produto Y foi elevada em 1,25 estrelas. Vide Figura 18.

Se um consumidor optar por adquirir um MP3 / MP4 *Player* apenas considerando sua reputação em um site de vendas online, sua decisão poderia ter sido influenciada por opiniões de pessoas que não necessariamente conhecem o produto. Ao invés de comprar a mercadoria Y, ela foi induzida a adquirir um item potencialmente pior, graças a ação voluntária ou não de certos indivíduos.

Figura 18 – *Rating* de produtos concorrentes após o recálculo



Fonte: elaborado pelo autor

Os resultados obtidos até aqui, entretanto, servem como análise das 270 *reviews* anotadas pelos voluntários dessa pesquisa, aproximadamente 3% do corpus do Buscapé. A fim de evitar que seja necessário anotar manualmente o restante das reviews, será apresentado, no próximo capítulo, a proposta de um estudo introdutório de *Machine Learning* para classificação automática das opiniões.

6 MODELO DE CLASSIFICAÇÃO DE *REVIEWS*

A obtenção de exemplos de *reviews* rotuladas entre Experientes e Outras propicia, portanto, a elaboração de um classificador supervisionado de *Machine Learning,* assim como foi proposto por diversas pesquisas na área de spams de opinião, como as citadas nesse trabalho e também por Bensouda, Fkihi e Faizi (2018) em sua revisão da literatura.

Por aprendizado supervisionado, Russel e Norvig (2013) entendem como um agente que observa um exemplo de entrada-saída, mapeando uma função que liga uma entrada até sua saída. Por isso, é essencial que se tenham exemplos rotulados. Como são apenas duas classes possíveis de *reviews*, Russel e Norvig (2013) atribuem a essa atividade, a nomenclatura de “classificação binária”.

Este capítulo descreve os métodos e ferramentas utilizados para a criação de um classificador de *reviews* preliminar. Por preliminar, entende-se que nem todas as características presentes no corpus foram exploradas, já que poderiam extrapolar o tempo disponível para desenvolvimento deste trabalho.

Vários modelos foram apresentados para o tratamento das correntes comportamentais e linguísticas descritas por Sandulescu e Ester (2015). O modelo computacional desta pesquisa segue a segunda vertente, relacionada apenas ao estudo de características textuais, visto que as contribuições realizadas pelos voluntários se basearam unicamente na leitura dos textos das opiniões, além dos nomes e categorias dos produtos. Pelo mesmo motivo, decidiu-se pela não exploração dos atributos comportamentais. Além do que, eles não se modificam conforme com a língua e já foram estudados em Li et al. (2015), Lim et al. (2010), Mukherjee, Liu e Glance (2012), entre outras publicações já descritas na seção 2.2.

Analisando apenas características textuais, técnicas de Processamento da Linguagem Natural (PLN) são naturalmente aplicáveis, dado que seu objetivo é justamente “adquirir informações a partir da linguagem escrita” (RUSSEL; NORVIG, 2013, p. 860, tradução nossa). Manning e Schutze (1999) explicam que o objetivo da ciência linguística é caracterizar e explicar observações textuais que nos cercam, como em conversas e na escrita.

PLN é uma disciplina da Inteligência Artificial, a qual analisa textos que ocorrem naturalmente e não fabricados especificamente para serem processados (LIDDY, 2001). Comprovando a importância das técnicas de PLN e ML, Bensouda, Fkihi e Faizi (2018) concluíram que os trabalhos mais relevantes na identificação de spams de opinião, fizeram uso dessa combinação.

Ao longo deste capítulo, são expostos os meios para extração de atributos dos textos das *reviews*. Em seguida, os algoritmos aplicados para a elaboração de classificadores são introduzidos. Com os resultados apresentados, avalia-se a eficácia da metodologia utilizada, comparando com modelos computacionais implementados por trabalhos relacionados.

6.1 PRÉ-PROCESSAMENTO

A fim de identificar atributos que representam as características da fração do corpus anotado em *reviews* Experientes e Outras, utilizou-se a biblioteca *Natural Language Toolkit* (NLTK), desenvolvida por Bird, Klein e Loper (2009) para a linguagem de programação Python. Ela foi escolhida por ser capaz de implementar diversas técnicas de PLN, além de possuir recursos que permitem trabalhar com a língua portuguesa (BIRD; KLEIN; LOPER, 2009).

De acordo com seus criadores, a origem da NLTK deu-se em 2001, “como parte de um curso de linguística na Universidade da Pensilvânia, Estados Unidos. De lá para cá, ela vem sendo adotada por cursos em dúzias de universidades” (BIRD; KLEIN; LOPER, 2009, prefácio, tradução nossa). Seus principais benefícios são:

- **Simplicidade:** Intuitivo, capaz de evitar a tediosa tarefa de processar dados linguísticos;
- **Consistência:** Uniforme, com estruturas de dados e interfaces consistentes, além de nomes de métodos dedutíveis;
- **Extensibilidade:** Novos módulos podem ser facilmente acomodados, os quais incluem implementações alternativas para abordar as mesmas tarefas; e
- **Modularidade:** Propicia componentes independentes sem a necessidade de entender a biblioteca inteira a fim de aplicá-los.

Antes de explorar a NLTK, entretanto, é necessário ter instalado um interpretador de comandos em Python. Bird, Klein e Loper (2009) propõem a utilização da própria linha de comando instalada juntamente com o Python[1]. Porém, para este trabalho, será utilizada ferramenta Jupyter Notebook[2], uma aplicação web que permite a criação e compartilhamento de documentos que contém códigos de diferentes linguagens de programação, equações e visualizações de dados (PROJECT JUPYTER, 2014).

Durante o processo de anotação, de acordo com o capítulo 4, percebeu-se que a linguagem encontrada em avaliações sobre produtos é informal, ou seja, repleta de gírias, abreviações, além de erros gramaticais. Duran, Nunes e Avanço (2015) estudaram o mesmo corpus do Buscapé de Hartmann et al. (2014) com o intuito de desenvolver uma ferramenta capaz de remover esses ruídos linguísticos por palavras mais comuns do vocabulário, ou seja, normalizando-as. Para a presente pesquisa, nenhuma normalização foi aplicada. Isso porque acredita-se que as expressões mais informais podem influenciar a classificação da *reviews*. Portanto, remover este atributo poderia significar uma perda na caracterização do corpus.

Pelo mesmo motivo, a *steaming*, técnica que retira afixos, deixando apenas os radicais da palavra (MANNING; SCHÜTZE, 1999), também não foi empregada. Pressupõe-se que as palavras “comprarei” e “comprei”, por exemplo, possam ter efeitos diferentes nas análises realizadas pelos anotadores, logo, aplicando *steaming* implica na possibilidade de descaracterizar a opinião. Enquanto “comprarei” representa uma ação a ser tomada, “comprei” é uma tarefa já realizada no passado e podem ser cruciais na tentativa de detectar se o usuário tem ou não experiência com a mercadoria.

Inicialmente, as *reviews* selecionadas por meio da metodologia descrita no capítulo 5 tiveram que passar por um pré-processamento, com o intuito de facilitar tarefas que venham a ser aplicadas em cima dos dados. Manning e Schütze (1999) explicam que dependendo do corpus, diferentes formas de formatação podem ser utilizadas. Contudo, cuidados devem ser tomados, principalmente quando é necessário identificar elementos específicos, como nomes próprios.

---

[1] Download disponível em: www.python.org.

[2] Download disponível em: www.jupyter.org.

O primeiro passo foi transformar todas as letras em minúsculas e remover a acentuação das mesmas. A Figura 10 (ver capítulo 4.2) demonstrou um exemplo de opinião extraída do corpus. Nela, verifica-se que o site do Buscapé estrutura o texto em três partes: a opinião propriamente dita, “o que gostei” e “o que não gostei”. Portanto, o segundo passo foi separar a *review* em três segmentos, removendo as expressões “o que gostei” e “o que não gostei” inseridas pelo Buscapé. O Quadro 4 apresenta o resultado da divisão da opinião disposta na Figura 10.

Outra técnica essencial utilizada em PLN é a tokenização. Normalmente, os textos são divididos em unidades chamadas de *tokens*, onde cada uma representa uma palavra, número ou sinal de pontuação (MANNING; SCHÜTZE, 1999). A biblioteca NLTK disponibiliza recursos prontos para implementação dessa técnica.

Quadro 4 – Divisão da opinião em três partes

Parte	Texto
1	Fui muito bem atendido. Valeu pelo esforço de vocês.
2	Estou muito feliz, mas comprar assim é tanto que complicado pois os tamanhos nem sempre batem com o que se espera.
3	Não tenho certeza, mas acho que o número 10 nas costas está meio torto.

Fonte: elaborado pelo autor

A tokenização reduziu o esforço para a extração de algumas características dos textos, como a quantidade de palavras, letras e pontuações. Isso porque tal método cria uma lista dos *tokens*, retirando espaços consecutivos e quebras de linha, comuns no ambiente online e que podem causar problemas se não tratados corretamente. Após a tokenização, faz-se importante remover as *stop words*.

Bird, Klein e Loper (2009), comentam que a NLTK possui corpora de *stop words* para 11 idiomas diferentes, incluindo o português. Segundo eles, *stop words* são palavras de alta frequência nos textos e que por vezes, é necessário filtrá-las antes de continuar o processamento. “*Stop words* são palavras de função que podem ser ignoradas em tarefas de recuperação de informações em abordagens orientadas a extração de palavras-chave sem afetar a eficiência dessa recuperação” (MANNING; SCHÜTZE, 1999, p. 533).

Utilizando como base o corpus do NLTK, a frase “valeu pelo esforço de vocês” presente no Quadro 4, após a remoção de *stop words* resulta em “valeu esforço vocês”. Percebe-se que as palavras “pelo” e “de” foram descartadas, por não

apresentarem conteúdo relevante para análise linguística. Ainda assim, o trecho destacado ainda apresenta sentido, já que as palavras-chave foram mantidas.

Com a finalização do pré-processamento aplicado nos textos das *reviews*, torna-se realizável a identificação de algumas características importantes do corpus, as quais serão cruciais para o sucesso do classificador de opiniões.

6.2 EXTRAÇÃO DE ATRIBUTOS

A maioria dos algoritmos de *Machine Learning* são desenvolvidos com o objetivo de selecionar os atributos mais apropriados e promissores para a tomada de decisão (WITTEN; FRANK, 2016). Em teoria, complementam Witten e Frank (2016, p. 288, tradução nossa), “nunca selecionariam atributos irrelevantes ou inúteis”. Porém, na prática, modelos de aprendizado como árvores de decisão, abordadas na seção 6.3, acabam sofrendo com propriedades que não podem ser aproveitadas, pois acabam levando em consideração apenas alguns exemplos de treinamento. Uma das melhores maneiras de selecionar atributos relevantes é através da escolha manual, baseado no entendimento profundo dos dados (WITTEN; FRANK, 2016).

Para a proposição do classificador de opiniões, foram selecionados 130 atributos linguísticos. Contudo, é provável que alguns sejam descartados para as técnicas de ML utilizadas. Endereçando este problema, foram propostas três categorias de atributos, separadas em dois tipos: individuais e de grupo. Com estes agrupamentos, os quais são explicados a seguir, pretende-se avaliar a importância de propriedades semelhantes do corpus, visando encontrar o melhor modelo computacional, ainda que em fase preliminar de criação.

6.2.1 Atributos individuais

Essa categoria de características se refere, principalmente, a atributos retirados individualmente das *reviews*, buscando selecionar propriedades únicas de cada opinião. A partir do Quadro 5, identificam-se cinco informações retiradas dos textos.

Além da categoria do produto, algumas propriedades estatísticas também foram selecionadas. Entre elas, estão o número de caracteres, quantidade de palavras e percentual de símbolos de pontuação, como ponto final, vírgula, ponto de exclamação, entre outros. Outro atributo retirado está relacionado às críticas feitas

aos produtos no texto da avaliação. De acordo com a seção 6.1, as *reviews* do Buscapé mantêm uma parte referente aos pontos negativos identificados pelos usuários, chamada “o que não gostei”. Pensou-se que seria importante indicar quantos caracteres esta fração da opinião contém, a fim de entender se as críticas podem induzir a anotação do usuário. Estes atributos foram inspirados no conjunto de 44 características extraídos por Costa, Benevenuto e Merschmann (2013).

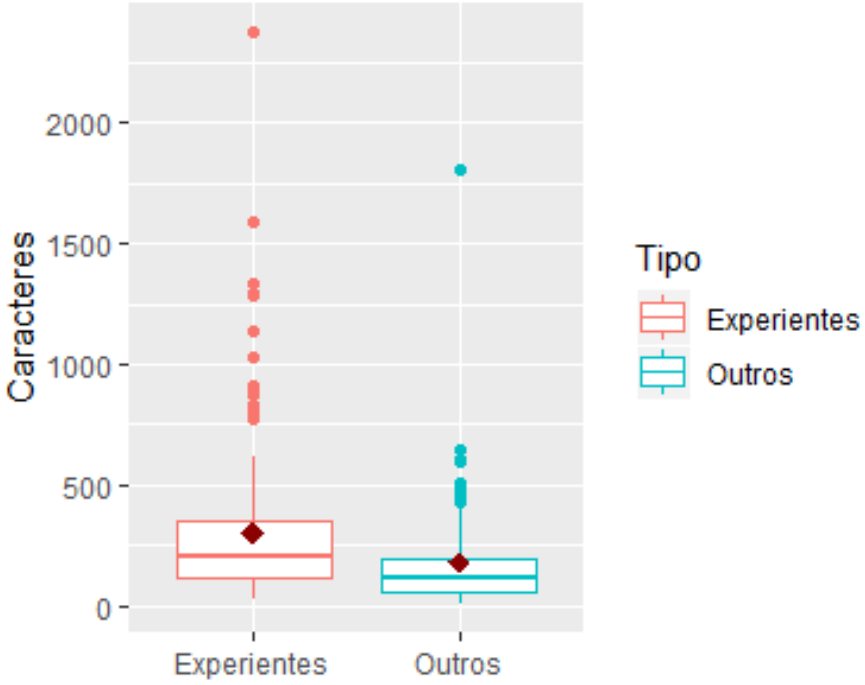
Quadro 5 – Atributos de opinião

Atributo	Descrição
Categoria do produto	Identificador da categoria do produto.
Caracteres da opinião	Quantidade de caracteres que a opinião possui.
Caracteres da crítica	Quantidade de caracteres digitados na seção “o que não gostei”.
Fator de pontuação	Percentual de pontuações presentes na opinião.
Palavras da opinião	Número de palavras da opinião.

Fonte: elaborado pelo autor

Pelo gráfico da Figura 19, é perceptível que o número de caracteres presentes nas opiniões de Experientes é maior do que os escritos em Outras. Em média, *reviews* de usuários que possuem experiência com o produto têm 300 caracteres (representado pelo losango vermelho), enquanto os que não têm, chegam a 179 letras, números e outros símbolos combinados. A mediana calculada é de 208 e 117 caracteres, respectivamente.

Figura 19 – Número de caracteres entre as duas classes de opiniões



Fonte: elaborado pelo autor

O mesmo comportamento é observado na quantidade de palavras. Para opiniões de Experientes, a média foi de 31 palavras, enquanto na classe Outros, 18,5. Isso indica a tendência de que avaliações maiores estão propensas a serem classificadas como Experientes. Já r*eviews* de tamanho menor acabam deixando mais dúvidas no processo de anotação.

Referente ao número de caracteres da crítica da opinião, também foi identificada uma grande diferença entre as duas classes de *reviews.* Para os Experientes, a média de caracteres foi de 63,6, enquanto para o tipo Outras, 15,9. Sem dúvidas, elencar pontos negativos do produto implica na maior aceitação de que o postador teve uma experiência genuína com o produto. O percentual de pontuações dentro do texto não sofreu alteração significante entre Experientes e Outras. Foram encontrados os valores de 3,64% e 2,88%, respectivamente.

Juntamente com os atributos de opinião, também foram selecionados mais características para compor o corpus de entrada para o modelo de ML, chamados de atributos de grupo.

6.2.2 Atributos de grupo

Ao contrário das propriedades anteriores, as quais se referiam somente ao texto da *review* individualmente, as próximas características possuem o objetivo de encontrar padrões comuns entre as classes, bem como para cada categoria de produto contida no corpus. Estes atributos foram inspirados na pesquisa de Ott et al. (2011). Para muitos, considerado o estado da arte em pesquisas na área de spams de opinião (MUKHERJEE et al., 2013).

Primeiramente, foram contados os unigramas e bigramas mais frequentes em cada classe de *review*. Russel e Norvig (2013) explicam que uma sequência de *n* símbolos, palavras, sílabas e até outras unidades de texto são chamadas de *n*-gramas, sendo os mais comuns: 1-grama (unigrama), 2-gramas (bigrama) e 3-gramas (trigrama). Para este trabalho, somente os unigramas e bigramas foram analisados, visto que, pela limitada quantidade de *reviews* no corpus, houve pouca repetição das mesmas três palavras em sequência.

Afirma-se que 270 é um número baixo de *reviews,* dado que este número é demasiadamente inferior aos dos corpus expostos na seção 3.1, inclusive

comparando-o com Costa, Benevenuto e Merschmann (2013), que ultrapassaram 7.000 opiniões em seu conjunto de opiniões.

A soma do número de ocorrências de cada termo foi proposta com o objetivo de detectar n-gramas capazes de generalizar os tipos de *review*, ou seja, expressões que quando encontradas, podem identificar se a mesma é Experiente ou Outras.

Foram selecionados os dez unigramas e dez bigramas mais frequentes em cada classe de opinião, os quais estão dispostos no Quadro 6. Também são descritos quantas vezes o termo foi encontrado dentro das *reviews*.

Quadro 6 – N-gramas mais frequentes por classe

N-grama	Experientes (nº de ocorrências)	Outras (nº de ocorrências)
Unigramas	bom (62) qualidade (49) excelente (41) uso (36) fácil (36) ótimo (34) preço (30) comprei (28) bonito (27) recomendo (26)	bom (26) qualidade (16) preço (14) excelente (13) ótimo (10) tudo (8) contra (8) benefício (8) legal (7) custo (7)
Bigramas	não tenho (17) custo benefício (17) assistência técnica (13) fácil manuseio (9) produto bom (9) bateria dura (8) excelente produto (8) poderia ser (8) boa qualidade (7) nada contra (7)	nada contra (6) não tenho (5) custo benefício (5) produto excelente (5) ótimo produto (4) tudo bom (3) tenho nada (3) quadro alumínio (3) melhor custo (3) dia dia (3)

Fonte: elaborado pelo autor

Verifica-se que, nos unigramas, há muitas palavras encontradas nos dois tipos de opinião, por exemplo, “bom”, “qualidade”, “excelente” e “preço”. O que chama a atenção, são os termos “uso”, “comprei” e “recomendo”, presentes apenas na lista de Experientes. Nas opiniões do tipo Outras, apesar de haver unigramas únicos, eles se repetiram poucas vezes, até pelo menor número de instâncias comparando com as *reviews* Experientes.

Já nos bigramas, destaca-se o termo “assistência técnica”, o qual apareceu em casos onde o postador da avaliação descreveu o processo de defeito do produto. Falar sobre características específicas do produto, segundo o Quadro 6, não

necessariamente pode ser classificado como Experientes, como alguém poderia pensar. “Bateria dura”, referindo-se a um celular ou notebook e “quadro alumínio”, identificando uma bicicleta, estão em classes opostas.

Para a montagem dos atributos, seguiu-se o seguinte método: para cada n-grama listado no Quadro 6 correspondente à classe da *review*, foi verificado se o mesmo estava contido no texto da opinião. Se sim, recebia o valor 1 (positivo), senão, 0 (negativo). Após esse procedimento, adicionou-se 17 atributos ao corpus, dado que alguns n-gramas se repetiram em ambas as classes.

Como exemplo, o n-grama “recomendo”, foi encontrado apenas em cerca de 13% das *reviews*. Ainda assim, não é garantido que esta palavra pode ser usada em grandes bases de dados como possível identificador de opiniões Experientes, pelo baixo número de *reviews* analisadas. Optou-se também por outra técnica de extração de atributos, chamada TF-IDF, sigla em inglês para *Term Frequency – Inverse Document Frequency*.

Ramos (2003, p. 972, tradução nossa) descreve que o TF-IDF

> Funciona pela determinação da frequência relativa das palavras em um documento específico comparado com a proporção inversa daquela palavra sobre todo o corpus de documentos. Intuitivamente, esse cálculo determina o quão relevante uma dada palavra é em um documento específico.

Manning e Schütze (1999) explicam que uma palavra que ocorre em todos os documentos, tem seu peso igual a 0 (mínimo) e se ela aparece somente em um documento, seu peso é 1 (máximo). Palavras como preposições, artigos e pronomes não possuem relevância, visto que são comuns na linguagem. Por isso, elas recebem baixos índices de TF-IDF (RAMOS, 2003). No cenário desta pesquisa, cada documento pode ser entendido como uma *review* ou conjunto delas.

Aizawa (2003) afirma que este método sinaliza relevância às palavras mais utilizadas em sistemas de Recuperação de Informação, do inglês, *Information Retrieval* (IR). A biblioteca NLTK também disponibiliza o cálculo de TF-IDF, escolhido para determinar o peso das palavras no corpus (BIRD; KLEIN; LOPER, 2009).

Diferentemente do cálculo de ocorrências dos n-gramas nas *reviews*, o intuito do cálculo do TF-IDF é identificar palavras que se destacam em um subconjunto de opiniões, desde que não apareçam com frequência no restante. Com base nesse conceito, propôs-se o agrupamento de *reviews* por categoria de produto, visando

assim, analisar os unigramas com maior relevância que pudessem identificar uma *review* Experiente ou Outras dentro de uma categoria.

Assim, tornou-se possível comparar, por exemplo, o documento das opiniões relacionadas a notebooks com os documentos das demais categorias, com o objetivo de identificar características específicas de notebooks. Esse procedimento foi realizado para 10 categorias diferentes, elencadas no Quadro 7. Apesar de totalizarem 53, apenas 10 categorias possuem mais de uma opinião etiquetada nas duas classes de *reviews*. Com apenas uma avaliação anotado como Experiente ou Outras, acredita-se que o cálculo de TF-IDF poderia ficar tendencioso, já que estaria sendo analisada a opinião de um único usuário.

No Quadro 7 são encontrados os cinco unigramas com maior índice de TF-IDF para cada categoria de produto e classe. Novamente, pelo baixo número de *reviews*, o TF-IDF não pôde ser calculado para bigramas ou até trigramas, dado que algumas categorias não conseguiram chegar a três termos com o índice maior que zero. Portanto, apenas unigramas foram analisados.

Quadro 7 – Unigramas com maior TF-IDF por categoria

(continua)

Categoria	Experientes	Outras
Aparelho de Pressão	qualidade preço tenho recomendo vizinhos	preço qualidade pessoas garantia cidade
Aparelho de Telefone (Fixo)	tenho telefone qualidade ligações eco	uso qualidade ótimo nenhum nada
Bicicleta	satisfeita pedal melhores elite crank	caloi brasil melhores crank brothers
Câmera Digital	recursos fácil máquina câmera uso	utilização recomendável positiva ótima muito
Celular e *Smartphone*	bateria câmera tela celular qualidade	tenho relação ótima nada gostei

Quadro 7 – Unigramas com maior TF-IDF por categoria

(conclusão)

Geladeira/Refrigerador	uso porta geladeira refrigerador meses	simplesmente qualidade lindo excelente eficiente
Jogos (Digitais)	jogo recomendo nada jogabilidade gráficos	tudo tocar guitarra grande gostei
Micro-ondas	grill micro-ondas cce pipoca função	especificações correspondem slim qualidade publicidade
Notebook	bateria aparelho excelente tela notebook	desempenho excelente designer busca bate
Perfume	perfume fixação ótima fragrância super	perfume embalagem preço nada quer

Fonte: elaborado pelo autor

O propósito desse trabalho não é discutir se o índice TF-IDF para cada n-grama é baixo ou alto, portanto, os resultados calculados não foram apresentados. Na categoria “celulares e *smartphones*”, por exemplo, é identificado um comportamento distinto entre os usuários que possuem experiência com a mercadoria em relação aos que não têm. Neste caso, em contrapartida com os atributos de grupo com base na ocorrência das expressões, abordar características do produto proporcionou separar as classes. Enquanto os experientes escreveram sobre a tela, bateria e câmera, os outros descreveram o produto em carácter genérico, utilizando adjetivos como ótima e gostei.

O processo para geração dos atributos com base no TF-IDF foi semelhante ao de ocorrências. Para cada *review*, foi verificado se a mesma continha algum unigrama listado no Quadro 7 para a mesma categoria, recebendo valores 0 (negativo) e 1 (positivo). Uma opinião de micro-ondas, por mais que tivesse a palavra “recomendo”, ainda assim recebia o valor 0, visto que o unigrama não é um dos termos com maior índice de TF-IDF dessa categoria.

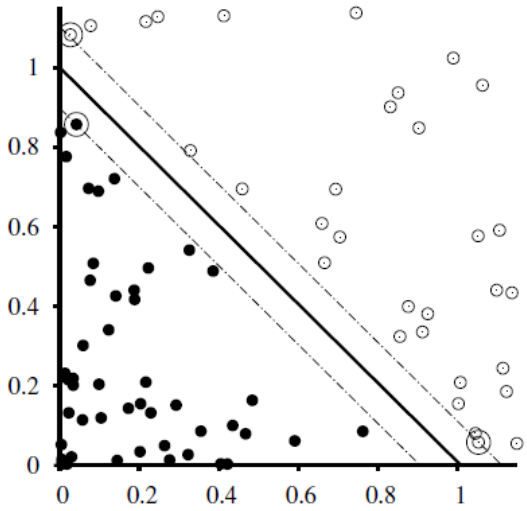
Ao todo, 98 atributos diferentes foram extraídos a partir desse processo, totalizando 130 em conjunto com as duas categorias de propriedades apresentadas anteriormente para a proposição do classificador de *reviews*. Eles servem como entrada para a predição da classe das opiniões. Na próxima seção, são descritos os algoritmos utilizados, bem como os resultados de eficácia do modelo de *Machine Learning*.

6.3 CLASSIFICADOR DE OPINIÕES

Por ser o único trabalho relacionado que avalia *reviews* em português, será seguida a metodologia de ML aplicada por Costa, Benevenuto e Merschmann (2013). Os autores utilizaram os algoritmos de *Support Vector Machine* (SVM) e *Random Forest* (RF), considerados o estado da arte em técnicas de classificação (COSTA; BENEVENUTO; MERSCHMANN, 2013).

O framework SVM é uma abordagem popular em aprendizagem supervisionada. Este algoritmo trabalha considerando que alguns exemplos podem ser mais importantes que outros e, levando isso em conta, é mais fácil chegar em generalizações (RUSSEL; NORVIG, 2013). A ideia do SVM é criar um plano ótimo de separação, ou seja, uma fronteira linear que separe os dados de treinamento. É como se cada exemplo fosse um ponto em um espaço N-dimensional (COSTA; BENEVENUTO; MERSCHMANN, 2013). De acordo com a Figura 20, verifica-se que há uma linha separando os exemplos pretos e brancos (classificação binária).

Figura 20 – Exemplo visual de SVM



Fonte: adaptado de Russel e Norvig (2013)

Russel e Norvig (2013) explicam que as linhas tracejadas representam a margem máxima calculada pelo algoritmo entre a linha de decisão e os exemplos mais próximos dela.

Já o *Random Forest* é um “classificador constituído de uma coleção de classificadores estruturados em árvore, [...] onde cada árvore vota na classe mais popular dado um *input x*” (BREIMAN, 2001, p. 6, tradução nossa). Segundo Costa, Benevenuto e Merschmann (2013, p. 36), o RF “constrói muitas árvores de decisão (floresta) e escolhe a classe com maior número de votos em todas as árvores da floresta”. Cada árvore depende de um vetor de exemplos independentes e aleatórios, porém com o mesmo número de *inputs*. Isso faz com que erros de generalização diminuam à medida que o número de árvores aumenta (BREIMAN, 2001).

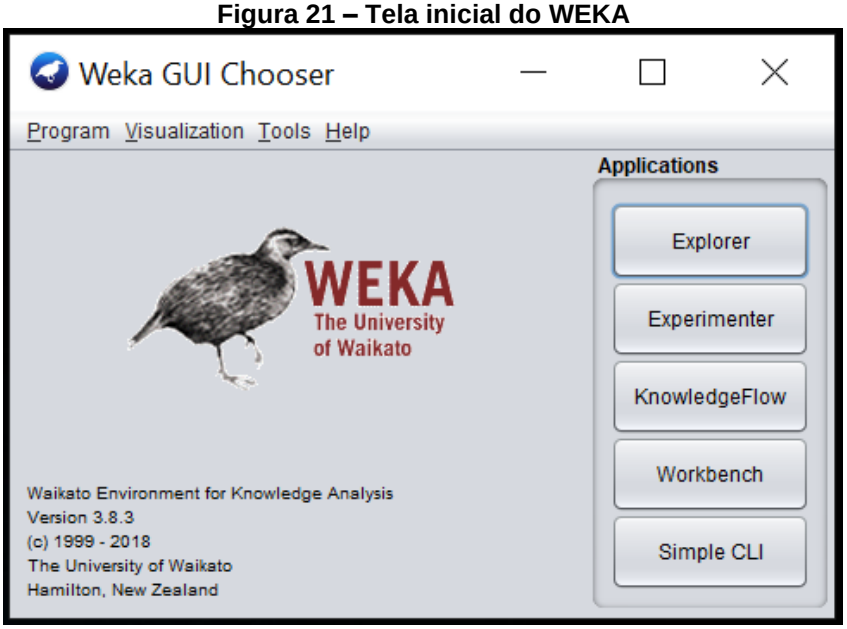
Ambos os algoritmos estão presentes na aplicação *Waikato Environment for Knowledge Analysis* (WEKA) (WITTEN; FRANK, 2016), que é utilizada para experimentação do SVM e RF neste trabalho, assim como foi realizado por Costa, Benevenuto e Merschmann (2013). Esta ferramenta foi criada para possibilitar a implementação de algoritmos de aprendizagem, juntamente com métodos para o pré e pós-processamento, avaliando os resultados de modelos em qualquer base de dados. Entre os recursos disponíveis, estão técnicas de regressão, classificação, associação e seleção de atributos (WITTEN; FRANK, 2016).

Para cada algoritmo, o WEKA (ilustrado na Figura 21) proporciona configurações padrão prontas para serem aplicadas. Entretanto, buscando o maior percentual de acerto na predição das classes das *reviews*, alguns parâmetros foram modificados. Para o SVM, utilizou-se o *kernel Radio Basis Function* (RBF), permitindo que os modelos gerados pela ferramenta consigam realizar separações mesmo com fronteiras complexas (COSTA; BENEVENUTO; MERSCHMANN, 2013). Juntamente com o *kernel*, outras duas configurações do SVM foram alteradas: *c* (custo) e *gama*. Nos melhores resultados obtidos, os valores desses dois parâmetros estavam em 10 e 0,1, respectivamente. Já em Costa, Benevenuto e Merschmann (2013), os critérios foram 100 para *c* e 1 para *gama.* Com relação ao RF, o parâmetro número de *features* foi ajustado para 1. Costa, Benevenuto e Merschmann (2013) propuseram o valor 7. Em todos os experimentos, os mesmos valores foram aplicados.

Para a medição do desempenho de predição da classe de cada *review*, utilizou-se o método *Cross Validation* 5 *fold.* Esta técnica funciona através da separação da

base de dados em cinco partes iguais. O modelo é gerado com base nos exemplos anotados nos 4 primeiros conjuntos e o quinta parte é utilizada para avaliar o classificador (COSTA; BENEVENUTO; MERSCHMANN, 2013). Esse processo é repetido 5 vezes, onde a cada iteração, um conjunto de validação diferente é escolhido (COSTA; BENEVENUTO; MERSCHMANN, 2013).

Figura 21 – Tela inicial do WEKA



Fonte: Witten e Frank (2016).

Para a montagem do classificador de ML, foi utilizada uma base de dados balanceada, ou seja, com o mesmo número de instâncias das duas classes possíveis. Levando em consideração que a quantidade final de *reviews* anotadas do tipo Outras foi 74, o mesmo número de opiniões Experientes foram selecionadas a partir do total de 192 disponíveis.

Essa seleção foi realizada com base em dois critérios. Primeiramente, somente as *reviews* cujas categorias de produtos analisados pelo índice de TF-IDF foram escolhidas. Assim, o número de avaliações caiu para 101. No segundo critério, mais 27 *reviews* foram descartadas das categorias de maneira proporcional. Dentro de cada categoria, a exclusão das avaliações foi feita de maneira aleatória.

Para cada algoritmo de ML, foram utilizados quatro conjuntos diferentes de atributos. Esta separação tem como finalidade avaliar o nível de relevância dos diferentes tipos de propriedades extraídas do corpus, conforme as explicações sobre os grupos de atributos descritos nas seções 6.2.1 e 6.2.2. Os agrupamentos propostos são:

- Conjunto A: apenas atributos individuais;
- Conjunto B: união do conjunto A e atributos de grupo com base no número de ocorrência dos termos;
- Conjunto C: união do conjunto A e atributos de grupo com base no TF-IDF;
- Conjunto D: a união de todos os atributos.

As únicas propriedades não retiradas nas experimentações foram as individuais. Isso porque algumas instâncias do corpus de entrada tiveram todos os atributos de grupo zerados, ou seja, nenhum unigrama ou bigrama foi encontrado nos textos. Logo, sempre mantendo o conjunto A nas gerações dos classificadores, estas *reviews* terão pelo menos cinco atributos que possam destacá-las das demais.

Os resultados, medidos em percentuais de acerto, podem ser visualizados na Tabela 10.

Tabela 10 – Percentual de acerto dos algoritmos SVM e RF

Técnica	Conjunto A (%)	Conjunto B (%)	Conjunto C (%)	Conjunto D (%)
SVM	68,92	76,35	76,35	77,70
RF	68,92	77,70	77,03	81,08

Fonte: elaborado pelo autor

O melhor classificador, portanto, foi gerado pelo *Random Forest,* no qual a taxa de acerto na predição da classe das opiniões foi de 81,08%. Com isso, em 120 das 148 *reviews*, o classificador previu corretamente seu tipo. A matriz de confusão pode ser consultada na Tabela 11. Os acertos são obtidos somando-se os valores da diagonal principal, já os erros, com base na diagonal secundária.

Tabela 11 – Matriz de confusão do classificador

		Classe real	
Classe predita		Outras	Experiente
	Outras	63	11
	Experiente	17	57

Fonte: elaborado pelo autor

Das 74 opiniões de Outras, o classificador previu sua classe corretamente em 63 (85,1%) dos casos. Já nas Experientes, acertou 57 (77%). Isso mostra, ainda que ligeiramente, maior facilidade em detectar possíveis spams postados por usuários que aparentam não ter experiência com o produto sendo avaliado.

Outro resultado que deve ser destacado é que somente com atributos individuais, o percentual não ultrapassou 68,92% nos dois algoritmos. Com a adição dos atributos de grupo, entretanto, o percentual subiu até 80%. Isso confirma que as extrações realizadas a partir do número de ocorrência dos termos no corpus, bem como do cálculo de TF-IDF, auxiliaram a tomada de decisão dos algoritmos.

Apesar dos resultados não estarem diretamente relacionados, os 81% de exatidão ultrapassam os 75% obtidos por Costa, Benevenuto e Merschmann (2013), assim como os 65% de Sandulescu e Ester (2015). Embora ainda fique atrás de trabalhos como o de Ott et al. (2011), o qual chegou a 87% de acerto, sendo considerado um dos melhores em detecção de spams.

Mesmo sendo apenas um estudo preliminar, foi possível detectar corretamente, um alto número de possíveis spams no corpus do Buscapé. Entretanto, a baixa quantidade de opiniões que foram consideradas para a geração do classificador pode ter facilitado a identificação, já que há menor variedade de opiniões.

7 CONCLUSÃO

A grande quantidade de *reviews* sobre produtos publicadas na Internet inviabiliza a possibilidade de possíveis consumidores lerem todas elas antes de tomarem decisões de compra. Por isso, muitas pesquisas visam sumarizar textos, a fim de apontar as características mais relevantes das opiniões disponíveis.

Ao se basearem em experiências de outras pessoas para compras online, pessoas passam a acreditar que quem posta opiniões em sites de venda realmente possui conhecimento sobre o produto que está avaliando. Entretanto, isso nem sempre acontece. Fabricantes e outros usuários podem utilizar-se de práticas fraudulentas, até mesmo inconscientemente, para aumentar a reputação de um alvo ou afetar a dos competidores.

Devido à maioria dos esforços para mitigar os impactos dessa prática, conhecida como spams de opinião, ocorrerem em textos na língua inglesa, esta pesquisa propôs a anotação de um corpus em português com o auxílio de voluntários. Buscou-se identificar manualmente opiniões que indicam que os usuários que a postaram não possuem experiência prévia com o produto. Logo, é factível acreditar que suas opiniões não são fundamentadas no uso da mercadoria, mas sim em achismos do que pode ser qualidade ou defeito.

Verificou-se que a obtenção de *reviews* anotadas manualmente, para que sejam utilizadas de exemplo em algoritmos de *Machine Learning*, é uma tarefa árdua. Isso é devido à necessidade de equilibrar quantidade e qualidade das etiquetações humanas. A metodologia aplicada nesta pesquisa buscou unir elementos de trabalhos relacionados, como *crowdsourcing* e análise de especialistas através do desenvolvimento do *website* Anoto-PT. Ferramenta essa, capaz de facilitar a aquisição de etiquetações, já que as contribuições dos voluntários são contabilizadas em tempo real, de qualquer lugar em que as pessoas convidadas estejam.

Entretanto, apenas uma pequena fração das *reviews* anotadas pôde ser considerada confiável, de acordo com o cálculo do coeficiente Kappa. Assim, é possível afirmar duas questões. A primeira está relacionada com a complexidade de humanos chegarem à um consenso sobre o que as opiniões podem significar. Já a

segunda vem da necessidade de aplicar treinamentos mais intensos aos anotadores, visando o ganho de conhecimentos, cada vez mais parecidos com o de especialistas.

Apesar do número de *reviews* efetivamente utilizadas para análise do impacto na reputação dos produtos ter sido reduzida, foi possível averiguar que há muitas opiniões passíveis de serem consideradas como spam. Tal constatação afirma a urgência de atacar esse problema que está induzindo consumidores a escolherem produtos que possam não atender suas necessidades.

Com carácter introdutório, também foi proposto um estudo preliminar na elaboração de um classificador automático de opiniões, utilizando os algoritmos SVM e *Random Forest*. Apenas propriedades linguísticas foram analisadas, dado que são as únicas que variam dependendo da língua sendo estudada, ao contrário dos atributos comportamentais, extensivamente abordados por outros autores. A extração de características dos textos através de n-gramas mostrou que, para o corpus utilizado na pesquisa, os mesmos foram relevantes para atingir significativas taxas de acertos na predição da classe das *reviews*.

Destaca-se que o desenvolvimento desta pesquisa proporcionou algumas contribuições para o campo científico. A partir corpus criado por Hartmann et al. (2014), sua aplicabilidade foi ampliada para o estudo de spams de opinião em português, pelo desenvolvimento de um *software* online para obtenção de anotações. Além disso, foram apresentadas ferramentas para exploração de técnicas de Processamento de Linguagem Natural e *Machine Learning*.

Algumas propostas de extensão desta pesquisa podem ser pensadas, como a obtenção de mais exemplos anotados, novas formas de treinamento de voluntários, utilização de atributos comportamentais complementares aos linguísticos e exploração de outros algoritmos de Inteligência Artificial. Técnicas para automatizar o processo de anotação do restante do corpus desempenham papel fundamental para possíveis aplicações comerciais, visto que o presente estudo abordou apenas a aprendizagem supervisionada de exemplos.

REFERÊNCIAS BIBLIOGRÁFICAS

AIZAWA, Akiko. An information-theoretic perspective of tf--idf measures. **Information Processing & Management**, [s. l.], v. 39, n. 1, p. 45–65, 2003. Disponível em: <https://ccc.inaoep.mx/~villasen/index_archivos/cursoTL/articulos/Aizawa-tf-idfMeasures.pdf>. Acesso em: 27 out. 2019.

AMAZON. **Mindhunter: o primeiro caçador de serial killers american**. 2019. Disponível em: <https://www.amazon.com.br/Mindhunter-primeiro-caçador-killers-americano-ebook/>. Acesso em: 3 mar. 2019.

APONTADOR. **Resultados para restaurante**. 2019. Disponível em: <https://www.apontador.com.br/local/search.html?q=restaurante&loc=>. Acesso em: 5 jun. 2019.

BBC. **Como a Amazon se transformou na empresa mais valiosa do mundo**. 2019. Disponível em: <https://g1.globo.com/economia/tecnologia/noticia/2019/01/09/como-a-amazon-se-transformou-na-empresa-mais-valiosa-do-mundo.ghtml>. Acesso em: 7 nov. 2019.

BENSOUDA, Nissrine; EL FKIHI, Sanaa; FAIZI, Rdouan. Opinion Spam Detection: A Review of the Literature. **Proceedings of the international conference on learning and optimization algorithms: theory and applications**, Rabat, mai. 2018.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python: analyzing text with the natural language toolkit**. [s.l.] : O'Reilly Media Inc., 2009. Disponível em: <http://www.datascienceassn.org/sites/default/files/Natural Language Processing with Python.pdf>. Acesso em: 26 out. 2019.

BISHOP, Todd. **Amazon files first-ever suit over fake product reviews, alleging sites sold fraudulent praise**. 2015. Disponível em: <https://www.geekwire.com/2015/amazon-files-first-ever-suit-over-fake-reviews-alleging-calif-man-sold-fraudulent-praise-for-products/>. Acesso em: 17 mar. 2019.

BREIMAN, Leo. Random forests. **Machine learning**, [s. l.], v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf>. Acesso em: 27 out. 2019.

BUSCAPÉ. **Smartphone Xiaomi Redmi Note 7 64GB**. 2019. Disponível em: <https://www.buscape.com.br/smartphone-xiaomi-redmi-note-7-64gb>. Acesso em: 7 nov. 2019.

CARDOSO, Emerson Freitas; ALMEIDA, Tiago A. Detecção Automática de Opiniões Falsas com base no Conteúdo das Mensagens. **Anais do 14th Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'17)**, Uberlândia, MG, p. 2–15, 2017. Disponível em: <http://www.dt.fee.unicamp.br/~tiago/papers/ERT-ENIAC17.pdf>. Acesso em: 3 mar. 2019.

COHEN, Jacob. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Nova Iorque, v. 20, n. 1, p. 37–46, 1960. Disponível em: <https://journals.sagepub.com/doi/10.1177/001316446002000104>. Acesso em: 19 mai. 2019.

CONDORI, Roque Enrique López; PARDO, Thiago Alexandre Salgueiro. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. **Expert Systems with Applications**, [s. l.], v. 78, p. 124–134, 2017. Disponível: <https://sites.google.com/icmc.usp.br/opinando/>. Acesso em: 7 mar. 2019.

COSTA, Helen; BENEVENUTO, Fabricio; MERSCHMANN, Luiz Henrique de Campos. **Detectando Avaliações Spam em uma Rede Social Baseada em Localização**. 2013. 71 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Ouro Preto, Ouro Preto, SP, 2013. Disponível em: <https://www.repositorio.ufop.br/handle/123456789/3413>. Acesso em: 16 mar. 2019.

DICIONÁRIO DA LÍNGUA PORTUGUESA MICHAELIS. **Dicionário Brasileiro da Língua Portuguesa**. 2019. Disponível em: <http://michaelis.uol.com.br/moderno-portugues/>. Acesso em: 14 maio. 2019.

DOSCIATTI, Mariza Miola; FERREIRA, Lohann Paterno Coutinho; PARAISO, Emerson Cabrera. Anotando um Corpus de Notícias para a Análise de Sentimentos: um Relato de Experiência. **Symposium in Information and Human Language Technology**, Natal, RN, p. 4-7, nov. 2015. Disponível em: <https://www.aclweb.org/anthology/W15-5616>. Acesso em: 20 abr. 2019.

DRURY, Brett et al. An Open Source Tool for Crowd-Sourcing the Manual Annotation of Texts. **Internacional conference on computacional processind of the portuguese language**, São Carlos, SP, 2014.

DURAN, Magali Sanches; NUNES, Maria das Graças Volpe; AVANÇO, Lucas. A normalizer for ugc in brazilian portuguese. **Proceedings os the workshop on noisy user-generated text**. Beijing. Disponível em: <https://bdpi.usp.br/item/002712653>. Acesso em: 26 out. 2019.

EUGENIO, Barbara Di; GLASS, Michael. The kappa statistic: A second look. **Computational linguistics**, [s. l.], v. 30, n. 1, p. 95–101, 2004. Disponível em: < https://www.eecis.udel.edu/~carberry/CIS-885/Papers/DiEugenio-Kappa-Second-Look.pdf>. Acesso em: 21 set. 2019.

FLEISS, Joseph L. Measuring nominal scale agreement among many raters. **Psychological bulletin**, [s. l.], v. 76, n. 5, p. 378, 1971.

FREITAS, LARISSA A. DE; VIEIRA, RENATA. Comparing portuguese opiniolexicons in feature-based sentiment analysis. **IJCLA**, [s. l.], v. 4, n. 1, p. 147-158, jun. 2013. Disponível em: <https://www.ijcla.org/2013-1/IJCLA-2013-1-pp-147-158-08-Comparing.pdf>. Acesso em: 18 mai. 2019.

HARTMANN, Nathan et al. A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words. In: LREC 2014, **Internacional Conference on Language Resources and Evaluation**, Reykjavik, Islândia, p. 3865-3871, mai. 2014. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/413_Paper.pdf>. Acesso em: 1 mai. 2019.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. **Internacional Conference on Knowledge Discovery and Data Mining**, Seattle, ago. 2004. Disponível em: <https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf>. Acesso em: 21 abr. 2019.

JINDAL, Nitin; LIU, Bing. Opinion spam and analysis. **International Conference on Web Search and Data Mining**, Palo Alto, Califórnia, fev. 2008. Disponível em: <https://www.cs.uic.edu/~liub/FBS/opinion-spam-WSDM-08.pdf>. Acesso em: 20 abr. 2019.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, mar. 1977. Disponível em: <https://www.dentalage.co.uk/wp-content/uploads/2014/09/landis_jr__koch_gg_1977_kappa_and_observer_agreement.pdf>. Acesso em: 15 mai. 2019.

LI, Fangtao Huang et al. Learning to identify review spam. **IJCAI**, 2011. Disponível em: <https://www.ijcai.org/Proceedings/11/Papers/414.pdf>. Acesso em: 14 mai. 2019.

LI, Huayi et al. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. **Internacional AAAI Conference on Web and Social Media**, [s. l.], p. 634-637, 2015. Disponível em: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/.../10461>. Acesso em: 12 mai. 2019.

LIDDY, Elizabeth D. Natural language processing. **Encyclopedia of Library and Information Science**, Nova Iorque, 2 ed, 2001. Disponível em: <https://surface.syr.edu/cnlp/11/>. Acesso em: 26 out. 2019.

LIM, Ee-Peng et al. Detecting product review spammers using rating behaviors. **Internacional Conference on Information and Knowledge Management**, Toronto, p. 939-948, out. 2010. Disponível em: <https://www.cs.uic.edu/~liub/publications/cikm-2010-final-spam.pdf>. Acesso em: 12 mai. 2019.

LIN, Yuming et al. Towards online anti-opinion spam: Spotting fake reviews from the review sequence. **Internacional Conference on Advances in Social Networks Analysis and Mining**, Pequim, p. 261-264, ago. 2014. Disponível em: <https://ieeexplore.ieee.org/document/6921594>. Acesso em: 21 abr. 2019.

LINGUATECA. **CETENFolha: Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo**. 2002. Disponível em: <https://www.linguateca.pt/cetenfolha/index_info.html>. Acesso em: 1 jun. 2019.

LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, [s. l.], v. 5, n. 1, p. 1–167, 2012.

LUCA, Michael; ZERVAS, Georgios. Fake it till you make it: Reputation, competition, and Yelp review fraud. **Management Science**, [s. l.], v. 62, n. 12, p. 3412–3427, 2016. Disponível em: <http://people.hbs.edu/mluca/FakeItTillYouMakeIt.pdf>. Acesso em: 28 abr. 2019.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge: MIT press, 1999. Disponível em: <https://www.cs.vassar.edu/~cs366/docs/Manning_Schuetze_StatisticalNLP.pdf>. Acesso em: 22 out. 2019.

MCHUGH, Mary L. Interrater reliability: the kappa statistic. **Biochemia medica: Biochemia medica**, Zagreb, v. 22, n. 3, p. 276–282, 2012. Disponível em: <https://hrcak.srce.hr/89395>. Acesso em: 26 mai. 2019.

MUKHERJEE, Arjun et al. What yelp fake review filter might be doing? **Internacional AAAI Conference on Weblogs and Social Media**, [s. l.], p. 409-418, 2013. Disponível em: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006>. Acesso em: 28 abr. 2019.

MUKHERJEE, Arjun; LIU, Bing; GLANCE, Natalie. Spotting fake reviewer groups in consumer reviews. **Internacional Conference on World Wide Web**, Lyon, abr. 2012. Disponíve em: <https://www.cs.uic.edu/~liub/publications/WWW-2012-group-spam-camera-final.pdf>. Acesso em: 21 abr. 2019.

MUKHERJEE, Arjun; VENKATARAMAN, Vivek. Opinion spam detection: An unsupervised approach using generative models. **Techincal Report, UH**, [s. l.], 2014. Disponível em: <https://pdfs.semanticscholar.org/b09b/a1d2e6b0437cd2de5a99beeff13491a1cd44.pdf>. Acesso em: 12 mai. 2019.

MURPHY, Rosie. **Local Consumer Review Survey**. 2018. Disponível em: <https://www.brightlocal.com/research/local-consumer-review-survey/#local-business-review-habits>. Acesso em: 17 mar. 2019.

OTT, Myle et al. Finding deceptive opinion spam by any stretch of the imagination. **Meeting of the Association for Computacional Linguistics**, Portland, p. 309-319, jun. 2011. Disponível em: <https://myleott.com/op_spamACL2011.pdf>. Acesso em: 2 mai. 2019.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, [s. l.], v. 2, n. 1-2, p. 1–135, 2008. Disponível em: <http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>. Acesso em: 21 abr. 2019.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Novo Hamburgo, RS: Feevale, 2013.

RALEIGH, Len. **The Yelp Purge of 2018 – Yelp's New Algorithm Filters More Reviews**. 2018. Disponível em: <https://www.telapost.com/yelp-purge-2018/>. Acesso em: 21 abr. 2019.

RAMOS, Juan. Using tf-idf to determine word relevance in document queries. **Proceedings of the first instructional conference on Machine Learning**, Nova Jérsei, 2003. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf>. Acesso em: 27 out. 2019.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência Artificial**. 3ª ed ed. Rio de Janeiro: Elsevier, 2013.

SANDULESCO, Vlad; ESTER, Martin. Detecting Singleton Review Spammers Using Semantic Similarity. **World Wide Web Conference**, Florença, mai. 2015. Disponível em: <http://www2015.wwwconference.org/documents/proceedings/companion/p971.pdf>. Acesso em: 25 mar. 2019.

SANTOS, Allisfrank Dos; JÚNIOR, Jorge Daniel Barros; CAMARGO, Heloisa de Arruda. Annotation of a Corpus of Tweets for Sentiment Analysis. **Internacional conference on computational processing of the portuguese language**, ago. 2018.

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. **Understanding machine learning:** From theory to algorithms. Nova Iorque: Cambridge university press, 2014.

SQUIRES, Lauren. Enregistering internet language. **Language in Society**, Nova Iorque, v. 39, n. 4, p. 457–492, 2010. Disponível em: <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/F8A79BB74879D022D911F3B818B727BF/S0047404510000412a.pdf/enregistering_internet_language.pdf>. Acesso em: 1 jun. 2019.

VIERA, Anthony J.; GARRETT, Joanne M. Understanding interobserver agreement: the kappa statistic. **Fam med**, [s. l.], v. 37, n. 5, p. 360–363, 2005. Disponível em: <http://www1.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf>

WANG, Aobo; HOANG, Cong Duy; KAN, Min-Yen. Perspectives on Crowdsourcing Annotations for Natural Language Processing. **Lang. Resour. Eval.**, Secaucus, NJ, Estados Unidos, v. 47, n. 1, p. 9–31, 2013. Disponível em: <http://dx.doi.org/10.1007/s10579-012-9176-1>. Acesso em: 1 set. 2019.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: Practical machine learning tools and techniques**. [s.l.] : Morgan Kaufmann, 2016.

YAO, Yao et al. Yelp's Review Filtering Algorithm. **SMU Data Science Review**, [s. l.], v. 1, n. 3, p. 1-33, 2018. Disponível em: <https://scholar.smu.edu/cgi/viewcontent.cgi?article=1029&context=datasciencereview>. Acesso em: 4 jun. 2019.

YUAN, Yuan et al. Interpretable and effective opinion spam detection via temporal patterns mining across websites. **IEEE Internacional Conference on Big Data**, Washington, p. 96-105, dez. 2016. Disponível em: <http://www.cse.lehigh.edu/~sxie/paper/bigdata16a.pdf>. Acesso em: 14 mai. 2019.

ZHANG, Rong et al. Exploiting shopping and reviewing behavior to re-score online evaluations. **Internacional Conference on World Wide Web**, Lyon, p. 649-650, abr. 2012. Disponível em: <https://dl.acm.org/citation.cfm?id=2188171&dl=ACM&coll=DL>. Acesso em: 14 mai. 2019.