QUANTIFICANDO A CONTRIBUIÇÃO DE EMOJIS E EMOTICONS PARA IDENTIFICAÇÃO DE POLARIDADE

HILDON LIMA DE PAULA

QUANTIFICANDO A CONTRIBUIÇÃO DE EMOJIS E EMOTICONS PARA IDENTIFICAÇÃO DE POLARIDADE

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Eduardo Freire Nakamura

Manaus

Junho de 2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

P324q

Paula, Hildon Eduardo Lima de
Quantificando a contribuição de emojis e emoticons para identificação de polaridade / Hildon Eduardo Lima de Paula. 2019
105 f.: il. color; 31 cm.

Orientador: Eduardo Freire Nakamura
Tese (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Mineração de Dados. 2. Aprendizagem de Máquina. 3. Identificação de Polaridade. 4. Emojis. 5. Ambientes Online. I. Nakamura, Eduardo Freire II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"QUANTIFICANDO A CONTRIBUIÇÃO DE EMOJIS E EMOTICONS PARA IDENTIFICAÇÃO DE POLARIDADE"

HILDON EDUARDO LIMA DE PAULA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Freire Nakamura - PRESIDENTE

Prof. Bruno Freitas Gadelha - MEMBRO INTERNO

Prof. Moisés Gomes de Carvalho - MEMBRO EXTERNO

Manaus, 31 de Maio de 2019

Dedico a toda minha família, amigos, e a todos que contribuíram direta ou indiretamente na conclusão desta dissertação.

Agradecimentos

Agradeço primeiramente a Deus, pois sem ele nada seria possível, à minha família, amigos e todas as pessoas que me ajudaram na conclusão deste trabalho. Agradecimento especial ao meu orientador, Prof. Dr. Eduardo Freire Nakamura, cuja orientação foi crucial em muitos momentos, e obrigado pela oportunidade de aprender e me desenvolver ainda mais. Obrigado ao Istituto de Computação (IComp) da UFAM, por prover toda a estrutura e conhecimentos necessários, desde a graduação até a dissertação de mestrado.

"Céus e terras passarão,
mas as minhas palavras jamais passarão"
(Mateus 24:35)

Resumo

Ambientes virtuais como lojas online de produtos e serviços (e.g. Amazon, Google Play, Booking) adotam uma estratégia colaborativa de avaliação e reputação onde os usuários classificam os produtos e serviços. A opinião do usuário representa o seu grau de satisfação em relação ao item avaliado. O conjunto de avaliações de um item é referencial de sua reputação/qualidade. Portanto, a identificação automática da satisfação do usuário em relação a um item, considerando sua avaliação textual, é uma ferramenta com potencial econômico singular. Neste contexto, com a popularização de emojis e emoticons, intensificada pelo uso de dispositivos móveis e seus aplicativos, os usuários adotam cada vez mais estes símbolos como parte do vocabulário utilizado para expressar opiniões e sentimentos. Neste trabalho, apresentamos uma avaliação quantitativa da representatividade de emojis/emoticons para a identificação de opinião e polaridade em ambientes online de avaliação colaborativa. A abordagem proposta quantifica o uso da técnica Bag of Words com SVM, Max Entropy e Naive Bayes para determinar o grau de satisfação do usuário em relação a um item, considerando: (1) palavras e emojis/emoticons; (2) apenas palavras; (3) apenas emojis/emoticons. Particularmente, para cenários específicos o uso de emojis/emoticons para a análise de sentimentos chega a ter uma eficácia de 0,92 com uso de emojis combinados com palavras, contra 0,81 quando utilizamos apenas as palavras, considerando a métrica F1.

Palavras-chave: Ambientes Virtuais, Mineração de Dados, Identificação de Polaridade, Emojis, Avaliações Online.

Abstract

Virtual environments, such as online stores (e.g. Amazon, Google Play, Booking), promote a collaborative strategy for reviewing products and services. The users' opinions represent their degree of satisfaction regarding the reviewed item. The set of reviews of an item serves as a reputation index. Hence, the automatic identification of the user satisfaction, regarding an item, based on his/her textual review, is a tool of great economic and strategic potential for enterprises. In this context, the growing adoption of emojis and emoticons, boosted by the mobile devices and their Apps, the users increasingly adopt such a vocabulary to express their opinion and sentiments. In this work, we present a quantitative assessment of the richness of emojis/emoticons to predict the users' opinion in product reviews in collaborative systems. Our proposal uses the Bag of Words with Support Vector Machine to predict the users' opinion in a online review, taking into account the use of: (1) only words; (2) words and emojis/emoticons; and (3) only emojis/emoticons. For certain scenarios, considering the F1 metric, the use of words and emojis results in an efficacy of 0.92 using words combined with emojis, compared to 0.81 when only words are used (traditional approach).

Keywords: Virtual Environments,E-commerce reviews, Data Minning, Polarity classification, Emojis.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Motivação e Contexto

Ambientes virtuais compreendem uma série de plataformas, como lojas virtuais, redes sociais, sites de comércio eletrônico e trocas, entre outros. Uma rede social é um conjunto de pessoas ou grupos com algum padrão de contato ou interação entre si, essas estão bem avançadas nos dias atuais [Haythornthwaite, 2005]. Uma forma de instanciar o conceito de redes sociais para a interação entre usuários a partir de uma ferramenta online é a Rede Social Online (RSO) [Mislove et al., 2007], onde a cada dia cresce a quantidade de usuários, seja pela facilidade de poder acessá-las em qualquer circunstância ou por *smartphones*, *tablets* ou PCs. Com isso, há uma enorme massa de dados sendo produzida pelos usuários, que expressam opiniões, sentimentos, notícias, entre outros. Desta forma, há possibilidades para estudos em relação ao comportamento dos usuários em ambientes virtuais [Baccianella et al., 2010; Miller, 1995; Pang and Lee, 2008; Wang et al., 2016].

Uma exemplo de rede social online é o Twitter, que permite aos usuários acessar recursos através de uma página web ou dispositivos móveis. O principal serviço que oferece é a possibilidade do usuário compartilhar um texto curto, de no máximo 140 caracteres (*tweet*), onde seus seguidores podem visualizar, favoritar e compartilhar essa postagem para seus seguidores pessoais.

Em junho de 2016, o número de usuários do Twitter ultrapassava a faixa de 310 milhões, já o número de tweets por dia, 500 milhões [Inc., 2016]. Por este motivo, o Twitter atrai um grande número de pesquisadores da área de processamento de linguagem natural [Aramaki et al., 2011]. Com isso, a mineração de dados do Twitter pode ser aplicada em diversas áreas para resolver os mais diversos problemas, tais como: previsão de epidemia, migração populacional e vigilância da opinião pública [Barnaghi

et al., 2016; Shah et al., 2016b].

Da mesma forma, em junho de 2018, o número de usuários do Instagram ultrapassou a marca de 1 bilhão de usuários [Inc., 2018]. Assim sendo, o Instagram é uma plataforma bastante visada pelos pesquisadores em diversas áreas de reconhecimento de imagem, análise de sentimentos [Gangrade et al., 2019], classificação de emoções [Illendula and Sheth, 2019], detecção de cyber bullying [Raisi and Huang, 2017].

Uma das áreas de destaque quando se estuda redes sociais online, é a opinião de usuários em relação a algum produto ou serviço. Por exemplo, grandes empresas e vendedores podem ter uma visão geral da opinião e sentimento dos usuários em relação aos seus produtos/serviços, dessa forma podem ajudar seus clientes a escolherem o produto mais adequado, melhorar alguns aspectos dos seus produtos ou até saber, com algum nível de significância, o quão interessante será o índice de vendas a partir das expectativas apresentadas pelos usuários [Wang et al., 2016]. Todas essas informações podem ser obtidas quase que em tempo real através de técnicas de análise de sentimento, classificação de polaridade e outras.

O problema de classificação de polaridade busca responder se um dado texto ou documento tem polaridade negativa, positiva ou neutro. Geralmente, pesquisas nessa área os dados utilizados são relacionados a alguma entidade ou produto e elas tentam entender e aprender qual o comportamento dos usuários em relação a eles. A classificação de polaridade tem muita utilidade em lojas virtuais como Google PlayStore, AppStore e Amazon. A partir das avaliações dos usuários sobre determinado App ou produto, os clientes podem decidir se irão comprá-lo ou não.

O reconhecimento de emoções e entidades tem sido uma área bastante estudada mesmo antes da existência dos microblogs como o Twitter. Já a análise de sentimentos é um conceito que encorpa uma série de processos como a extração, classificação, classificação subjetiva, sumarização de opinião e detecção de spams [Wang, 2010]. Para conseguir executar essas atividades, a área de análise de sentimento lida com inúmeros desafios.

Para a pesquisa em torno da análise de sentimentos, é necessário definir os conceitos de *opinião*, *subjetividade* e *emoção*. Todavia, essa tarefa não é trivial, pois um usuário do Twitter, por exemplo, pode escrever um *tweet* que possa carregar valor negativo positivo ou expressar notícias, valores ou até mesmo emoções específicas (raiva, desapontamento, etc). Todas essas definições precisam ser representadas matematicamente para que sirvam de entrada para suas respectivas atividades. O sucesso da análise de sentimento depende diretamente da capacidade de extrair características necessárias para cada conceito, além de executar a atividade correspondente [Serrano-Guerrero et al., 2015a].

Desde os tempos das antigas civilizações, o ser humano já tinha a necessidade de representar coisas, seja sua história, avisos, ou perigos. Desta forma, desenhos foram a primeira maneira de representação escrita pela humanidade. E dentro do contexto mais atual (década de 70-80), os emoticons eram frequentemente utilizados para reforçar o humor, e melhorar a semântica e o entedimento de contexto entre mensagens de texto. E na década de 90, Shigetaka Kurita decidiu apresentar uma evolução dos emoticons, com o intuíto de deixar os emoticons mais atrativos e bem humorados, que se chamavam emojis. Uma vantagem deste tipo de representação é que qualquer usuário, independente da lingua ou nação, consegue entender, pois são imagens.

Assim sendo, uma maneira muito simples de representar emoções e sentimentos, é através dos *emoticons* e *emojis* [Paiva, 2016]. Emoticons é a representação pictorial de uma face humana e suas emoções a partir de símbolos de pontuação e caracteres da tabela ASCII. Eles introduzem um tom mais casual e informativo aos textos dos usuários que o utilizam. Os *emojis* são representações mais figurativas que os *emoticons* e conseguem expressar mais emoções e de maneira específica.

Segundo uma das maiores redes sociais atualmente, o Facebook, 60 milhões de emojis são postados todos os dias, e 6 bilhões são enviados através do Messenger (App de mensagens do Facebook) [Burge, 2018]. E isso não se restringe apenas ao Facebook, naturalmente as pessoas tendem a utilizar emojis ao trocar mensagens via WhatsApp. Desta forma, é possível observar que a quantidade de usuários que utiliza *emoticons* e *emojis* em ambientes virtuais, lojas virtuais, e redes sociais é muito alta. Quanto mais casual o ambiente, maior a tendência de uso. Sendo assim, torna-se possível classificar a polaridade de textos e sentenças somente com o uso de *emoticons* e *emojis*, utilizando apenas características extraídas deles como entrada dos algoritmos de classificação, ou então utilizá-los como complemento às palavras do textos, visando a melhor classificação de polaridade de textos. Uma vez que utilizando apenas os *emojis*, a complexidade de execução dos algoritmos de classificação tende a ser menor, por que existem menos *emojis* e *emoticons* que palavras em muitos alfabetos.

1.2 Objetivos

Este trabalho tem o objetivo de demonstrar o ganho (na métrica F1) com o uso de emojis/emoticons para o problema de classificação de polaridade em ambientes virtuais online. Os objetivos específicos são:

1. Demonstrar que o método descrito nesta metodologia apresenta uma taxa de

acerto próxima ou superior ao estado da arte;

2. Demonstrar que o método apresentado identifica e classifica os conteúdos independente do domínio (ambiente de contexto) e que os emojis tem uma contribuição relevante neste processo;

3. Demonstrar que utilizando somente os emojis de um texto também é possível fazer a classificação com resultados próximos ou melhor do que fazendo uso somente das palavras.

1.3 Principais Contribuições

A principal contribuição deste trabalho é o método para quantificar a importância dos emojis/emoticons na identificação de polaridade e classificação de textos a partir de dados disponíveis na web, tais como avaliações de apps e comentários em redes sociais. Além disso, a pesquisa contribui com o estudo e experimentação da utilização de diferentes modelos de aprendizagem de máquina:

1. ***Survey* sobre as técnicas de análise de sentimento e detecção de polaridades**. Permite conhecer e ampliar percepções acerca das pesquisas que tratam de detecção de polaridade em relação ao estado da arte e identificar novas oportunidades nos estudos existentes que servem como pontos de partida para a pesquisa.

2. **Classificação e detecção de polaridade**: Busca demonstrar a implementação do método de identificação de polaridade, bem como a utilização da abordagem de aprendizagem de máquina com os dados coletados de ambientes virtuais online no intuito de quantificar a importância de emojis/emoticons nesta tarefa.

3. **Base de dados de avaliações da Google PlayStore**: Base de dados com mais de 1 milhão de avaliações documentadas e separadas. A base de dados será explicada a frente em detalhes.

1.4 Organização do Documento

O Capítulo 2 apresenta os conceitos necessários para uma melhor compreensão dos aspectos gerais do método proposto, bem como uma revisão da literatura relacionada à área de abrangência deste trabalho. Além disso, apresentar-se-á pontos importantes

sobre análise de sentimentos e identificação de polaridade, além dos diversos cenários aos quais foram aplicados.

O Capítulo 3 apresenta uma síntese dos principais trabalhos relacionados a área de análise de sentimentos e identificação de polaridade, e como os mesmos tratavam os emojis/emoticons e os resultados obtidos pelos mesmos. Também será apresentado o estado da arte nesta área, as abordagens, e métricas utilizadas para a validação de seus resultados.

O Capítulo 4 apresenta a abordagem proposta por este trabalho, bem como todas as fases que a compreendem. Será apresentado como foram obtidos os dados, as etapas de pré-processamentos utilizadas, e os métodos utilizados para a classificação de polaridade.

O Capítulo 5 demonstra como os experimentos foram realizados, os cenários abordados e as técnicas de extração de características utilizadas, e como foram organizados os experimentos.

No Capítulo 6 é apresentada a base de dados utilizadas nos experimentos, a maneira como foi construída, as ferramentas que foram utilizadas para a coleta, e uma análise mais detalhada dos dados coletados, número de classes, campos, entre outros.

O Capítulo 7 demonstra os experimentos e resultados obtidos com o uso dos algoritmos de classificação Naive Bayes, Max Entropy, Random Forest e SVM respectivamente. Neles foram apresentados os gráficos de comparação das métricas de aprendizagem para cada algoritmo e as matrizes de confusão.

O Capítulo 8 apresenta as conclusões finais que podem ser inferidas a partir dos resultados obtidos nos capítulos anteoriores. Além disso, são apresentadas as limitações da abordagem proposta, e os possíveis próximos passos da pesquisa.

Capítulo 2

Fundamentação Teórica

Nesse capítulo é apresentado um resumo dos principais conceitos e soluções recentes (de 2003 a 2019) para o problema de detecção e classificação de textos em redes sociais. O conjunto de sete trabalhos revisados abordam técnicas para identificar informações relevantes no Twitter por meio de agrupamentos de dados, com técnicas não-supervisionadas de aprendizagem de máquina, supervisionadas, e híbridas.

Os principais conceitos presentes na área de análise de sentimento (mineração de opinião, análise subjetiva, detecção de polaridade e outros) são utilizados no geral como sinônimos. Porém, suas origens não são as mesmas, e muitos autores consideram como coisas diferentes. Essa seção tem como objetivo, apresentar os conceitos necessários para o entendimento dá área de análise de sentimentos e detecção de polaridade.

2.1 Conceitos Principais

Uma opinião pode ser definida como de sentimento positivo ou negativo, subjetivo, emoção, de revisão sobre uma entidade (pessoa, produto, evento, tópico ou organização) ou sobre algum aspecto da opinião [Barnaghi et al., 2016].

Dessa forma, uma opinião pode ser definida matematicamente como uma tupla-5 $(e_j, a_{jk}, SO_{ijkl}, h_i, t_l)$ onde e_j é a entidade alvo e a_{jk} é o k-ésimo aspecto/característica da entidade e_j, SO_{ijkl} é o valor sentimental da opinião do usuário h_i no aspecto a_{jk} da entidade e_j no tempo t_l [Serrano-Guerrero et al., 2015a].

Há diversas maneiras de se classificar opiniões, porém, podemos destacar duas categorias, sendo elas: Regular e comparativa. A maioria das opiniões são regulares, e por sua vez podem ser subdivididas em diretas e indiretas. As diretas expressam algum sentimento/ideia em relação a alguma entidade ou aspecto, enquanto as indiretas expressam a opinião sobre uma entidade ou aspecto com base nos seus feitos em outras

entidades/aspectos. As opiniões comparativas demonstram as semelhanças ou pontos em comum entre as entidades ou aspectos [Jindal and Liu, 2006a,b; Liu, 2012]. Alguns autores ainda classificam as opiniões como implícitas ou explícitas, dependendo se as sentenças expressam opiniões subjetivas ou objetivas [Li, 2013].

Além de sentimento e opinião, existem dois conceitos próximos, subjetividade e emoção. De acordo com alguns autores, uma expressão subjetiva é aquela em que o sujeito expressa algum sentimento, opinião, ou valor pessoal [Li, 2013], todavia, não implica que toda sentença subjetiva tenha que expressar algum sentimento.

A diferença entre sentenças objetivas e subjetivas é que geralmente a objetiva expressa alguma informação factual do mundo, enquanto a subjetiva expressa algum valor pessoal, opinião ou sentimento. Na maioria das vezes, a subjetividade envolve sentimentos quando se está lidando com afeto, julgamento, apreciação, especulação, entre outros. Por outro lado, uma emoção pode ser vista como uma expressão de nossos próprios sentimentos e pensamentos subjetivos. Emoções são muito próximas dos sentimentos, de fato, a maneira de medir uma opinião está diretamente ligada à intensidade de certas emoções, como a raiva, surpresa, tristeza, medo, amor ou alegria. Um exemplo que pode ser dado é: "Amo a minha casa", onde o orador expressa objetivamente seu amor por sua casa [Serrano-Guerrero et al., 2015b].

2.2 Principais Desafios

Há muitas tarefas ligadas à análise de sentimentos. Algumas delas estão relacionadas e é difícil separá-las claramente pois compartilham muitos aspectos. As mais importantes são:

1. **Classificação do Sentimento**: Também conhecido como orientação do sentimento, orientação semântica ou polaridade do sentimento [Yu et al., 2013]. Essa tarefa se baseia na ideia de que um documento, ou sentença expressa uma opinião sobre uma entidade (ou tema) de um orador e tenta medir o sentimento desse orador pela entidade. Portanto, consiste principalmente em classificar as opiniões em positivas, negativas e neutras [Shah et al., 2016b]. Apesar de parecer simples, é uma tarefa complexa, ainda mais quando o pesquisador deseja abranger múltiplos domínios e línguas [Hussien et al., 2016]. Essa tarefa está relacionada à predição do nível de sentimento. Diferentes escalas podem ser utilizadas para medir uma opinião, por exemplo, a maioria dos autores utiliza [-1,1] onde -1 indica opinião negativa, e 1 positiva [Pak and Paroubek, 2010]. Outros autores classificam como [-1,0,1] onde 0 indica sentimento neutro [Shah et al.,

2016b]. Alguns também classificam de 1 a 5. Porém, independente das escalas que são utilizadas para classificar, podemos tratá-los como problema de detecção de polaridade com até 5 rótulos.

2. **Classificação Subjetiva**: O principal objetivo dessa tarefa é detectar se a sentença é subjetiva ou não. Uma frase objetiva expressa uma informação factual, enquanto uma sentença subjetiva pode expressar outros tipos de informações pessoais como: opiniões, avaliações, emoções e crenças. Além disso, sentenças subjetivas também podem ou não expressar sentimentos negativos. Essa tarefa pode ser vista como uma etapa que precede a classificação de sentimentos. Uma boa classificação de subjetividade pode garantir uma ótima acurácia no classificador de sentimentos. É considerada também uma tarefa mais difícil do que distinguir a sentença entre os sentimentos positivos, negativos e neutros [Montejo-Ráez et al., 2014].

3. **Sumarização de Opinião**: É uma tarefa focada principalmente em extrair características de uma entidade compartilhada por um ou mais documentos (ou sentenças) e identificar opiniões sobre ela [Selvan and Moh, 2015]. É a detecção de opinião em si, e não dos sentimentos.

4. **Classificação da polaridade**: É uma tarefa focada principalmente em extrair características de uma entidade compartilhada por um ou mais documentos (ou sentenças) e a polaridade relacionado a ela [Meng et al., 2012; Selvan and Moh, 2015]. Esses trabalhos podem ser divididos em relação ao número de documentos (pode ser uma sentença, ou várias). Geralmente estes trabalhos classificam os documentos em positivo, negativo e neutro.

5. **Detecção de Sarcasmo e Ironia**: Consiste na detecção de declarações que contém tom irônico ou sarcástico. Essa é uma das tarefas mais complexas e desafiadoras, visto que, há ausência de acordo entre pesquisadores sobre como a irônia ou sarcasmo podem ser definidos. Alguns trabalhos usam uma abordagem semelhante a classificação de sentimentos [Bharti et al., 2015; Bouazizi and Ohtsuki, 2016], se numa sentença os componentes da mesma exprimem muitos sentimentos inversos (positivos e negativos, ou, tristeza e alegria) então é detectado o sarcasmo, mas a acurácia máxima atualmente é de 83,1% com uma precisão de 91,1% [Bouazizi and Ohtsuki, 2016].

6. **Análise de Sentimento baseado em *emoticons***: Esta atividade pode ser vista como um sub-desafio da análise de sentimento, onde a partir de *emoticons*

presentes nas sentenças é possível classificá-las em sentimentos felicidade, tristeza, raiva, entre outros. É um desafio relativamente novo, que vem crescendo em pesquisas, principalmente nos domínios de linguagens que utilizam muitos símbolos, como chinês, coreano, japônes e linguagem arábica [Hussien et al., 2016].

7. **Outros**: Além das atividades mencionadas anteriormente, outras tarefas relacionadas à análise de sentimento podem ser destacadas, por exemplo, análise de gênero ou detecção de autoria, que tenta determinar o gênero ou a pessoa que escreveu um texto / opinião ou detecção de spam de opinião. Procurar detectar opiniões ou revisões que contenham conteúdos não confiáveis (Fake news), divulgadas para distorcer a opinião pública em relação a pessoas, empresas ou produtos.

2.3 Emojis e Emoticons

É comum, ao ocorrer eventos importantes, usuários expressarem seus sentimentos em relação a esses, mostrando como eles afetaram seu estado de humor. Para tal, diversos usuários tendem a utilizar emoticons, uma representação pictorial de uma expressão facial construída com números, letras e pontuações (e.g., =D, =/). Emoticons são frequentemente utilizados para reforçar o temperamento de um estado de humor, e podem melhorar ou até mudar a interpretação de um texto simples [Goncalves et al., 2013]. Os emoticons são construídos unicamente por caracteres do teclado padrão, o que facilita o seu uso em usuários de máquinas *desktop* e *notebooks*.

Durante os anos 90, no Japão, Shigetaka Kurita inventou uma nova forma de representação de sentimentos, com o intuito de facilitar o uso de emoticons e deixar as conversas com um tom mais descontraído, os emojis [Nakano, 2016]. Ao invés de usar apenas símbolos e pontuações, os emojis são códigos *unicodes* pré-definidos para a melhor representação de figuras (e.g., ☺, ☹). No entanto, os dispositivos base para os emojis na época eram os *pagers*, que tinham uma utilidade mais profissional, e por isso não se popularizou tanto. Porém, com o crescimento das redes sociais e a popularização dos *smartphones* e *tablets*, os emojis renasceram com a ideia de deixar mais fácil e descontraída as conversas e a representação dos sentimentos dos usuários. Em dispositivos móveis (*smartphones* e *tablets*), os teclados apresentam os emojis como se fossem teclas normais, facilitando assim o uso por parte dos usuários (diferente dos computadores onde temos teclados apenas com letras e símbolos).

Com o surgimento e a popularização da Internet, ocorreram diversas outras evoluções, como o surgimento das lojas virtuais, redes sociais, e diversos outros ambientes

de colaboração entre pessoas. As lojas virtuais ajudaram a internacionalizar o mundo das vendas, e troca de produtos entre usuários. E após a compra, ou contratação de algum serviço, normalmente o usuário tem a possibilidade de escrever uma avaliação sobre o produto e/ou sobre o vendedor do produto. Isso abre uma gama de possibilidades de estudo, uma vez que os usuários expressam suas opiniões e sentimentos a cerca do produto obtido, e dependendo da loja, ou ambiente, os usuários tendem a fazer uso de emojis e emoticons para expressar melhor o humor, e as emoções que ele está sentindo, além de dar um tom mais casual e informativo.

Dependendo do ambiente, quanto mais casual e informativo, mais os usuários tendem a usar os emojis. Uma rede social é composta por um conjunto de pessoas ou grupos com algum padrão de contato ou iteração entre si [Wives, 2013]. Os padrões de amizade entre esses indivíduos, relações e os casamentos entre famílias são exemplos de redes sociais que foram estudadas [Newman, 2003]. O conceito de redes sociais começou a evoluir à partir do trabalho de Wellman et al. [1996]. Segundo Wellman et al. [1996], uma rede social adota um mesmo suporte que as redes de computadores, pois interliga os usuários assim como as máquinas, formando tais redes. A partir dessa definição surgem os conceitos como: Redes sociais por computadores e Comunicação mediada por computadores.

Desta forma, os usuários tendem a expressar-se muito mais com o uso de emojis em redes sociais, e o número de usuários só aumenta a cada instante. A quantidade de dados gerados pelas redes sociais é muito vasta, o que abre margem para estudos em várias áreas, inclusive na classificação de polaridade com o uso de emojis. Na Figura 2.1. Nela, podemos identificar a polaridade dos tweets sobre o produto "Galaxy S10" utilizando os emojis.

Além do Twitter, podemos utilizar avaliações de lojas virtuais, como por exemplo a PlayStore da Google. Na Figura 2.2 dois usuários avaliaram um jogo na loja virtual, com o uso de emojis e emoticons.

Em ambientes de lojas virtuais, as avaliações já estão rotuladas pela sua própria nota. Nas redes sociais seria necessário um trabalho de rotulação manual dos textos.

Uma coisa interessante que pode ser observada, é que se optarmos por utilizar apenas os emojis dos textos, não dependeríamos da linguagem para a classificação, o que tornaria o classificador abrangente, pois a maioria das abordagens utiliza apenas uma linguagem específica (geralmente o inglês).

Dessa forma, abre a possibilidade para estudos a partir dos dados que contêm *emoticons*/*emojis* produzidos pelas redes sociais e ambientes virtuais. Uma vez que a análise de símbolos tende a ser mais simples que a de palavras, pois existem menos emoticons/emojis que palavras, eles independem de naturalidade da linguagem, e têm

um valor de representatividade de sentimentos maior que apenas as palavras em si.

2.4 Representação de texto

Algoritmos de aprendizagem de máquina não interpretam diretamente documentos digitais, sendo preciso transformar estes dados em documentos facilmente processáveis computacionalmente, uma representação que compacte o seu conteúdo [Goncalves et al., 2013].

2.4.1 Bag-of-Words

Este é o modelo de representação mais utilizado por algoritmos de classificação de texto. Este modelo transcreve a cadeia de caracteres de um documento em um vetor de palavras, guardando em cada posição além da própria palavra, também a sua frequência. Ele ignora toda a estrutura do documento, assim como a pontuação e a ordem das palavras. Por ser simples, ele não consegue identificar bem a semântica, ou seja, uma palavra que é usada com dois significados diferentes seriam tratadas iguais e teriam a mesma frequência. Ainda assim, este modelo por ser simples, tem a vantagem de ser processado rapidamente em relação a outras abordagens, o que é uma economia muito grande quando se tem uma base de dados com uma grande quantidade de texto.

No modelo de bag-of-words qualquer tipo de texto, palavra, ou documento é visto como um grupo de elementos. A contexto da aplicação, este elemento pode ser uma palavra simples, um n-grama, uma estrutura composto, JSON, XML, entre outros.

2.4.2 N-Gramas

O N-Gram é um vetor onde cada elemento é uma sequência de n palavras consecutivas, por exemplo se $n = 2$, cada posição do vetor tem duas palavras consecutivas do documento. Este modelo possibilita guardar a ordem das palavras, desta forma, consegue conservar o contexto de suas utilizações, o que é uma vantagem em relação ao bag-of-words. Por exemplo, duas palavras, dependendo de sua ordenação, podem dar um significado diferente para a frase: “Grande mulher” e “Mulher grande”. Neste exemplo a primeira expressão remete ao caráter, a personalidade, e o segundo remete a característica física da mulher.

2.5 Abordagens baseadas em Dicionários Léxicos

Abordagens léxicas se baseiam numa coleção de termos conhecidos dos sentimentos, frases e até idiomas, desenvolvida para gêneros comuns de comunicação [Wilson et al., 2005], também para estruturas mais complexas como ontologias [Kontopoulos et al., 2013], ou dicionários semânticos já mensurado o conteúdo em relação a orientação das palavras e termos em relação aos sentimentos.

Esse tipo de abordagem pode ser dividida em duas: Abordagem baseada em dicionário, ou baseado em *Corpus*. A primeira é geralmente baseada no uso de um conjunto inicial de termos (*Seeds*) que normalmente são coletados e anotados de forma manual. Esse conjunto cresce pesquisando os sinônimos e antônimos de um dicionário. Um exemplo desse dicionário pode ser a WordNet [Miller, 1995], que foi usado para desenvolver uma ferramenta chamada SentiWordNet [Baccianella et al., 2010]. A principal desvantagem desse tipo de abordagem é a incapacidade de domínio e orientações específicas do contexto. Todavia, soluções assim podem ser relevantes dependendo do problema, caso as sentenças não precisem necessariamente depender do contexto.

As baseadas em Corpus, por sua vez, surgem com o objetivo de fornecer dicionários relacionados a um domínio específico. Estes dicionários são gerados com um conjunto de termos de opinião *seed* que cresce através da busca de palavras-meios de utilização de técnicas estatísticas ou semânticas, o corpus é um corpo grande do texto da língua natural usado para acumular estatísticas no texto natural da língua.

Métodos baseados sobre estatísticas tais como Análise Semântica Latente (LSA) [Deerwester et al., 1990], ou simplesmente a frequência de ocorrência das palavras (*Term Frequency*) dentro de uma coleção de documentos podem ser utilizados [Pardal and Lopes, 2011]. Por outro lado, há métodos como o uso de sinônimos e antônimos ou relacionamentos de dicionário [Miller, 1995]. Alguns autores utilizam um dicionário léxico também como entrada para algumas etapas de caracterização de sentenças nas técnicas de aprendizagem de máquina [Bharti et al., 2015].

De acordo com Cambria et al. [2013], a análise de sentimento pode ser considerada um problema de PLN restrito, onde só é necessário compreender os sentimentos positivos ou sentimentos negativos relativos a cada sentença e/ou às entidades visadas ou tópicos. No entanto, apesar de ser um problema restrito, todos os trabalhos nesse campo, bem como todos os trabalhos em Recuperação de Informação, sempre enfrentam problemas com PLNs não resolvidos (manipulação de negação, reconhecimento de entidade nomeada, desambiguação de sentido de palavra, e etc) que são essenciais para detectar dispositivos literários como ironia ou sarcasmo [Reyes et al., 2012], e consequentemente, encontrar e avaliar sentimentos.

Um dos principais aspectos com que o PLN tem de lidar é com os diferentes níveis de análise. Dependendo se o alvo de estudo é: um texto curto ou documento, uma ou mais sentenças interligadas, uma ou várias entidades ou aspectos dessas entidades; Diferentes tarefas PLN e Análise de Sentimento podem ser realizadas. Assim, é necessário distinguir três níveis de análise que determina as diferentes tarefas da Análise do Sentimento:

1. Nível do documento;
2. Nível de sentença;
3. Nível de entidade/aspecto.

O nível de documento considera que um documento é uma opinião sobre um aspecto. Esse nível está associado à tarefa denominada *document-level sentiment classification* [Liu, 2012]. Entretanto, se um documento apresenta várias sentenças tratando de diferentes aspectos ou entidades, então um nível de sentença é mais adequado. O nível de sentença está relacionado com a tarefa conhecida como classificação subjetiva (do inglês, *subjetive classification*), a qual distingue sentenças que expressam informações factuais (objetivas) das sentenças que expressam visões subjetivas e opiniões pessoais. Finalmente, quando informações mais precisas são necessárias, então surge o nível de entidade/aspecto. É o nível mais refinado, considera que um alvo expressa uma opinião positiva ou negativa. Esse último nível é possivelmente o mais complexo, porque é necessário extrair com alta precisão muitos recursos como: datas ou intervalos de tempo, as diferentes características/aspectos e entidades, juntamente com as relações entre elas, as opiniões e as suas características, entre outros. Esse nível está estreitamente relacionado com características como *Opinion Mining* e *Opinion Summarization* [Ojokoh and Kayode, 2012]. Wang et al. [2005] demonstram uma abordagem semântica para classificação de polaridade de avaliações de produtos online. A abordagem faz uso POS-tagging para distinguir adjetivos e advérbios nas sentenças, e o algoritmo de classificação Naive Bayes junto com algumas regras heurísticas. Essa abordagem obteve em média 68% de acurácia. Bhargava et al. [2016] apresentam um modelo híbrido para computar perfis de vendedores confiáveis para e-commerce. O objetivo era comparar os vendedores classificados como confiáveis com o que é apresentado nos próprios sites. Primeiramente ele extraia palavras dos comentários de cada vendedor utilizando um dicionário de sentimentos e POS-tagging, isso buscava ignorar palavras pouco significativas para o modelo. Em seguida, com as sentenças filtradas apenas com palavras interessantes para classificação de sentimento, foi aplicado um modelo semi-supervisionado de classificação, que dava um score para cada review. Ao final era

feita a comparação do valor calculado para cada vendedor, com o seu respectivo valor no site.

2.6 Algoritmos de Aprendizagem

Os algoritmos de aprendizagem podem ser divididos em três principais tipos: supervisionado, não supervisionado e híbridos (combinação dos dois).

2.6.1 Algoritmos supervisionados

Os algoritmos supervisionados são aqueles onde há um agente supervisor que "ensina" o algoritmo, ou seja, entrega dados rotulados, de forma que este consiga mapear as características da entrada e relacioná-las ao valor esperado (rótulo). Durante a sessão de treinamento, o modelo ajustará suas variáveis para mapear as entradas e assim tentar prever as saídas correspondentes. Problemas de aprendizagem supervisionados são classificados em problemas de "regressão" e "classificação".

A diferença básica entre eles, é que os problemas de regressão estão tentando prever resultados de variáveis contínuas, ou seja, mapear a entrada para alguma função contínua. Já os de classificação estão tentando prever variáveis discretas, ou seja, tentam mapear características da entrada em classes ou categorias.

O maior desafio em utilizar este tipo de algoritmo se dá pela dificuldade em conseguir dados rotulados, pois quando se quer dados de redes sociais, por exemplo, é necessário fazer um processo de rotulação manual por humanos para se obter um certo grau de assertividade/confiança nos rótulos. Alguns exemplos de algoritmos são: SVM, Naive Bayes, e Max Entropy.

2.6.1.1 Support Vector Machines

As Máquinas de Vetores de Suporte (SVM) constituem uma técnica de aprendizagem de máquina embasada pela teoria de aprendizado estatístico desenvolvida por Vapnik [1995]. Essa técnica estabelece uma série de princípios que devem ser seguidos para classificadores obterem boa generalização, que se define por prever corretamente a classe de novos dados independente do domínio onde o aprendizado ocorreu.

O SVM pode ser utilizado para problemas de classificação com duas ou mais classes, e cada exemplo, ou dado, é representado por um vetor de características. Cada característica, também chamada de atributo, expressa um determinado aspecto do dado.

As SVMs podem ser lineares ou polinomiais (n classes). Para o linear, que posteriormente a formulação abrange problemas com mais de duas classes, funciona encontrando um ou mais hiperplanos de separação mais otimizada possível, que consigam classificar bem os documentos (dados/exemplo) e ainda assim mantém a capacidade de generalização do modelo. Abaixo um exemplo de representação de dados linear com um hiperplano de separação:

As SVMs lineares são eficazes na classificação de conjuntos de dados linearmente separáveis, ou que tenham uma distribuição aproximadamente linear, sendo que as lineares de margens suaves têm alguns mecanismos que toleram a presença de ruídos e outliers. Todavia, há casos em que os lineares não são capazes de dividir satisfatoriamente os dados de treinamento por um hiperplano.

As SVMs trabalham com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como entradas, para um novo espaço de maior dimensão, denominado espaço de características. Para mapear os dados, é necessária uma função, denominada função Kernel. Seja a função $Y : X- > S$, onde X é as entradas e Y denota o espaço de características. A escolha apropriada de Y faz com que o conjunto de treinamento mapeado em S possa ser separado por uma SVM linear.

2.6.1.2 Naive Bayes

Naive Bayes é uma técnica de classificação probabilística que assume que todas as características são condicionalmente independentes. Um exemplo de aplicação de Naive Bayes é a classificação de um documento em duas classes (*Spam* e não *Spam*). Uma representação muito simples, chamada o *bag-of-words*, serve para ignorar a ordem das palavras e apenas contar o número de vezes que cada palavra ocorre. Suponha que haja palavras D na linguagem. Então um documento pode ser representado como um p-vetor de contagens (um histograma de frequência de palavras). Seja $X = k$ significa que a palavra ocorre exatamente k vezes, para $k = 0 : K - 1$; Por simplicidade, vamos dizer que esta palavra tem contagem k. (Se a palavra ocorrer mais do que $K - 1$ vezes em um documento, apenas trataremos como se tivesse ocorrido $K - 1$ vezes; Aqui K é um limite superior escolhido pelo usuário).

Além do Naive Bayes comum, também existe o Naive Bayes Multinomial, onde são tratados vários vetores de características que representam as frequências que certos eventos foram gerados por uma multinomial $(p_1, ..., p_n)$ onde p_i é a probabilidade do evento i ocorrer (ou K caso seja multiclasse). Um vetor de características $\mathbf{x} = (x_1, \ldots, x_n)x = (x_1, \ldots, x_n)$ é então um histograma, com x_i contando o número de

vezes que o evento i foi observado em uma instância em particular. Este evento modelo tipicamente usado para classificação de documentos, com eventos representando a ocorrência das palavras em um único documento (como uma *bag-of-words*).

Se um dado valor de classe e recurso nunca ocorrer juntos nos dados de treinamento, então a estimativa de probabilidade baseada em frequência será zero. Isso é problemático porque eliminará todas as informações nas outras probabilidades quando elas forem multiplicadas. Portanto, muitas vezes é desejável incorporar uma correção de pequena amostra, chamada Pseudo Count [Bellemare et al., 2016], em todas as estimativas de probabilidade de modo que nenhuma probabilidade seja definida como sendo exatamente zero. Esta maneira de regularizar Bayes ingênuo é chamado de suavização Laplace quando o Pseudo Count é um, e suavização de Lidstone no caso geral.

Rennie et al. [2003] discute problemas com a premissa multinomial no contexto da classificação de documentos e possíveis maneiras de aliviar esses problemas, incluindo o uso de pesos TF-IDF (Term Frequency Inverted Document Frequency) em vez de frequências de termo bruto e normalização do comprimento do documento, para produzir um classificador Bayes ingênuo que é competitivo com as SVMs em termos de generalização.

2.6.1.3 Max Entropy

A técnica Max Entropy [Nigam et al., 1999] é um método que não assume interdependência entre os atributos e é feito o mínimo de restrições possíveis para estimar as probabilidades. As restrições expressam algum tipo de relacionamento entre os atributos e classes, e derivam do conjunto de treino. A distribuição de probabilidade que melhor satisfaz as restrições é aquela com maior entropia.

Max Entropy é uma técnica geral para estimar distribuições de probabilidade a partir dos dados. O excesso de equitação (princípio da Max Entropy) é que quando nada é conhecido, a distribuição deveria ser tão uniforme quanto possível, isto é, tem a Max Entropy. Dados de treinamento rotulados são utilizados para derivar um conjunto de restrições para o modelo caracterizar as expectativas específicas de classe para a distribuição. Características são representadas como valores esperados de "atributos" para qualquer função real, por exemplo. O algoritmo de escalonamento melhorado iterativo encontra a máxima distribuição de entropia que é consistente com o dados das restrições [Nigam et al., 1999].

Quando o problema tem muitas restrições são necessárias rigorosas técnicas para encontrar a solução ideal. Csiszar descreve bem várias técnicas úteis de Max Entropy [Csiszar, 1996]. A ferramenta utilizada neste trabalho [Pedregosa et al., 2011] utiliza

regressão logística para sua implementação, e seu parâmetro de execução foi apenas $C = 1e5$ para a utilização de todas as bases.

2.6.1.4 Random Forest

O Random Forest é um algoritmo de aprendizado de máquina flexível e fácil de usar que produz, mesmo sem ajuste de hyper-parameter, um ótimo resultado na maioria das vezes [Liaw and Wiener, 2002b]. É também um dos algoritmos mais utilizados, porque é simples e pode ser usado para tarefas de classificação e regressão. No Random Forest é fácil visualizar a importância relativa que atribui aos recursos de entrada [Tin Kam Ho, 1998].

A Random Forest também é considerada um algoritmo muito prático e fácil de usar, porque os hiperparâmetros padrão geralmente produzem um bom resultado de previsão. O número de hiperparâmetros também não é tão alto e eles são fáceis de entender.

Um dos grandes problemas no aprendizado de máquina é o Overfitting, mas na maioria das vezes isso não será fácil para um classificador de floresta aleatório. Isso porque, se houver árvores suficientes na floresta, o classificador não supera o modelo [Tin Kam Ho, 1998].

A principal limitação da Random Forest é que um grande número de árvores pode tornar o algoritmo lento e ineficaz para previsões em tempo real [Liaw and Wiener, 2002a]. Em geral, esses algoritmos são rápidos de treinar, porém são lentos para criar previsões depois de treinados. Uma previsão mais acurada requer mais árvores, o que resulta em um modelo mais lento. Na maioria das aplicações do mundo real, o algoritmo de floresta aleatória (random forest) é rápido o suficiente, todavia, podem haver situações em que o desempenho relacionado ao tempo de execução é importante e outras abordagens seriam preferidas.

E, claro, a Random Forest é uma ferramenta de modelagem preditiva e não uma ferramenta descritiva. Isso significa que se você estiver procurando por uma descrição dos relacionamentos em seus dados, outras abordagens seriam preferidas.

2.6.2 Algoritmos não supervisionados

Os não supervisionados, por sua vez, permitem abordar problemas com pouca ou nenhuma ideia dos resultados que se deseja obter. Neles, a base de dados não é rotulada, desta forma, não há supervisor para treinar o modelo. Neles procuramos agrupar os dados de maneira a traçar algum tipo de padrão para diferenciá-los.

O principal interesse do aprendizado não supervisionado é descobrir a organização dos possíveis padrões existentes nos dados através de clusters (agrupamentos) consistentes. Desta forma, é possível obter conclusões úteis a partir de similaridades e diferenças entres estes padrões.

Estes algoritmos também podem ser usados para reduzir o número de dimensões em um conjunto de dados para concentrar somente nos atributos mais úteis, ou para detectar tendências. Alguns algoritmos não-supervisionados: Algoritmos de Clustering (ex K-means), Hierárquico, Algoritmo Cocktail Party, e Singular-Value Decomposition.

2.6.2.1 Principal Component Analysis

O PCA é um algoritmo baseado na ideia de separar os dados a partir de características específicas ou frequentes em grandes volumes de dados [Abdi and Williams, 2010]. Este algoritmo provê redução de dimensionalidade. É bastante utilizando quando se tem uma grande quantidade de características, provavelmente muito correlacionadas entre si, e os modelos podem facilmente serem sobreajustados em um grande conjunto de dados.

PCA calcula a projeção dos dados em algum vetor que maximize a variância dos dados e perca a menor quantidade de informação possível. Surpreendentemente, estes vetores são os autovetores da matriz de correlação das características de um conjunto de dados.

2.6.2.2 K-means

K-means é um método de quantização vetorial, originalmente do processamento de sinais, que é popular para análise de cluster em mineração de dados. k-means clustering visa particionar n observações em k clusters nos quais cada observação pertence ao cluster com a média mais próxima, servindo como um protótipo do cluster. Isso resulta em um particionamento do espaço de dados nas células de Voronoi[Voronoi, 1908].

O problema é computacionalmente difícil (NP-difícil); no entanto, algoritmos heurísticos eficientes convergem rapidamente para um ótimo local. Estes são geralmente semelhantes ao algoritmo de maximização da expectativa para misturas de distribuições Gaussianas por meio de uma abordagem de refinamento iterativo empregada pela modelagem de mistura de k-médias e gaussianas. Ambos usam centros de cluster para modelar os dados; no entanto, o K-means clustering tende a encontrar clusters de extensão espacial comparável, enquanto o mecanismo de maximização de expectativas permite que clusters tenham formas diferentes. Uma coisa que vale a pena ressaltar, é que o K-means só clusteriza nativamente dados numéricos, assim sendo, para proces-

sar cadeia de caracteres é necessário fazer algum tipo de conversão ou transformação dos dados para tal. Pode-se analisar os termos mais utilizados em cada cluster, para assim traçar perfis e diferenciar os elementos. Ao usar Bag of Words como entrada do K-means, cada uma das suas sentenças é representada em um espaço dimensional alto de comprimento igual ao vocabulário. Para representar isto em 2D, é necessário reduzir a dimensão, por exemplo, usando o PCA com dois componentes.

2.7 Comentários finais do capítulo

Neste capítulo foram apresentados os principais conceitos na área de detecção e classificação de polaridade em textos, os termos, desafios e algoritmos desta área. Além disso, foi apresentado um conceito de opinião e como os trabalhos desta área o tratam em seus estudos. Neste trabalho o desafio a ser focado é o de detecção de polaridade, esta detecção se dará em avaliações on-line coletadas a partir da loja de aplicativos da Google, o que será abordado a frente.

Foi apresentado também um breve histórico dos emoticons e emojis, de onde surgiram, e sua utilização atualmente.

O uso dos emojis tem crescido constantemente com o advento de inúmeras redes sociais, e a facilidade que os teclados de smartphones provêm para a sua utilização. Boa parte de redes sociais, como Twitter, Facebook, e Instagram tem seus conteúdos com muita utilização. E não poderia ser de outra forma também em ambientes de lojas online, onde os usuários conseguem avaliar produtos que são comprados, ou aplicativos, e onde usuários conseguem interagir entre si ao avaliá-los.

Também foram apresentados formas de representar textos (Bag of words e N-gramas), para transformá-los em entradas mais facilmente processadas pelos algoritmos de aprendizagem. Foram apresentados tipos de algoritmos de aprendizagem, que se diferenciam em supervisionados (quando os dados estão rotulados de alguma forma) e não supervisionados (quando não há agente ou rótulo que ajude o algoritmo a aprender). Os principais algoritmos supervisionados são o SVM, Naive Bayes, Max Entropy e Random Forest. Estes quatro algoritmos foram avaliados neste trabalho, por serem os mais utilizados em outros trabalhos e pelo fato de a base de dados utilizada ja estar rotulada.

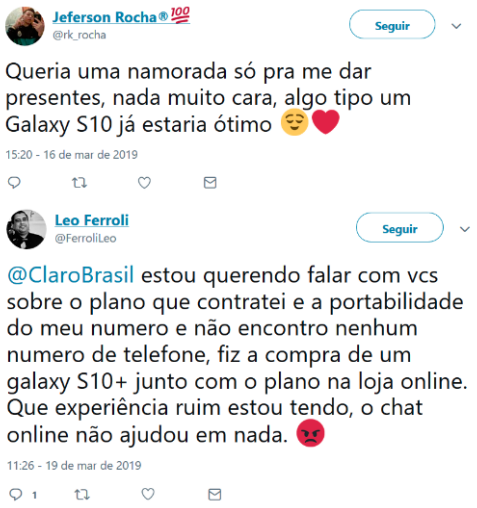


Figura 2.1: Exemplos de tweets com emojis.

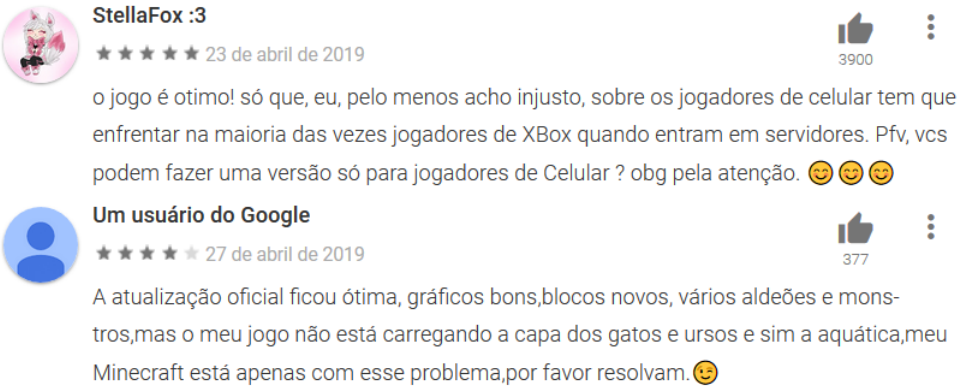


Figura 2.2: Exemplos de avaliações com emojis da Google PlayStore.

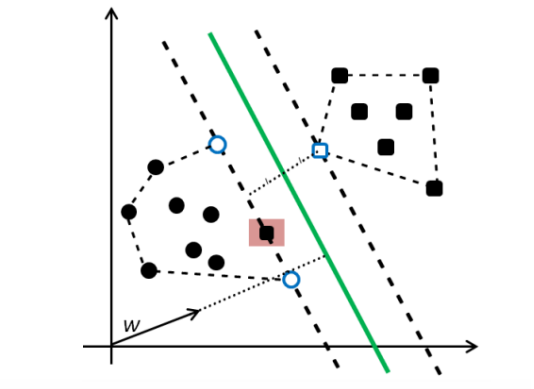


Figura 2.3: Espaço de características linear.

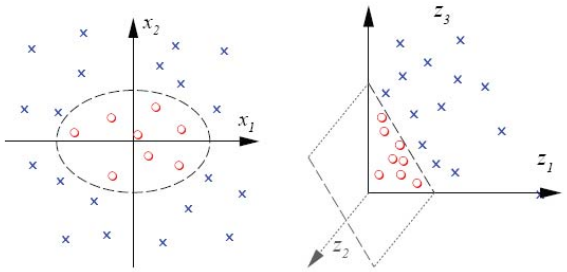


Figura 2.4: Espaço de características não-linear.

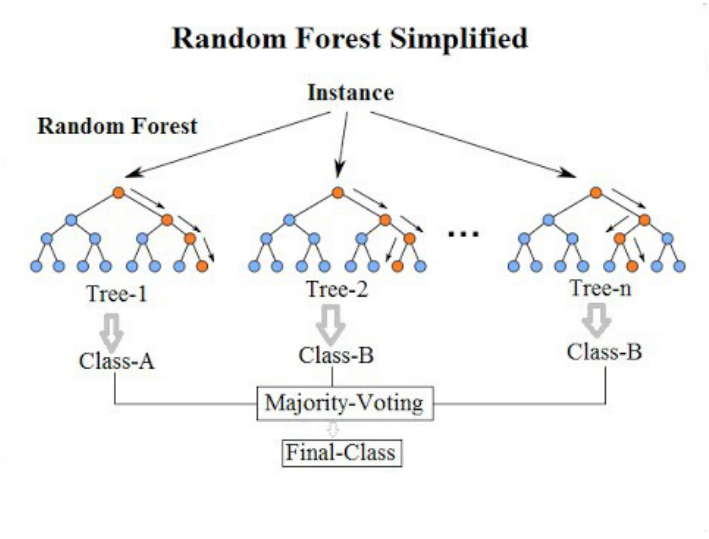


Figura 2.5: Random Forest simplificado [Breiman, 2001].

Capítulo 3

Trabalhos Relacionados

Nesse capítulo é apresentado um resumo dos principais conceitos e soluções recentes (2003 a 2019) para o problema de detecção e classificação de textos em ambientes online. O conjunto de trabalhos revisados abordam técnicas para identificar informações relevantes em ambientes online por meio de agrupamentos de dados com técnicas não-supervisionadas de aprendizagem de máquina, supervisionadas, e híbridas.

Os principais conceitos presentes na área de análise de sentimento (Mineração de Opinião, análise subjetiva, detecção de polaridade e outros) são utilizados no geral como sinônimos. Porém, suas origens não são as mesmas, e muitos autores consideram como coisas diferentes. Essa seção tem como objetivo, apresentar os conceitos necessários para o entendimento da área de análise de sentimentos e identificação de polaridade.

Com relação ao uso de *emojis* e *emoticons*, há diversos estudos relatando o impacto de seus usos na área de Ciências Sociais. Ip [2012] investigou o impacto dos *emoticons* na interpretação de afeto em mensagens instantâneas. Ela concluiu que o uso de *emoticons* ajuda os interlocutores a transmitir suas emoções durante a conversa on-line. Derks et al. [2007] investigaram a influência do contexto social sobre o uso de emoticons na comunicação na Internet.

Vários trabalhos tentam mostrar as diferentes técnicas aplicadas à análise de sentimentos. A maioria agrupa os trabalhos do ponto de vista de aplicações ou desafios nessa área de análise de sentimento [Pang and Lee, 2008]. Podemos agrupar os principais trabalhos da área de acordo com as técnicas utilizadas, como uma árvore, segundo Serrano-Guerrero et al. [2015b]. A Figura 3.1 apresenta as principais técnicas utilizadas em análise de sentimentos:

3.1 Abordagens de Aprendizagem de Máquina

O sucesso das técnicas estão diretamente relacionadas à extração do conjunto de características apropriada usadas na detecção de sentimentos. Nesta tarefa, técnicas de Processamento de Linguagem Natural (PLN) tem um papel importante, pois algumas das principais características mais utilizadas são, por exemplo [Medhat et al., 2014]:

1. Termos (Palavras ou n-gramas) e suas frequências;
2. Parte de informação de fala (Part-of-Speech), adjetivos são importantes nesta tarefa, mas substantivos também podem ter significância;
3. Negações podem inverter o sentido de qualquer sentença;
4. Dependências sintáticas (árvore de análise - *Tree parsing*) podem determinar o significado da sentença.

As principais técnicas supervisionadas utilizadas nos trabalhos são: Máquinas Vetoriais de Suporte (*Support Vector Machines* – SVM), *Naive Bayes*, *Maximum Entropy* e *Random Forest* [Barnaghi et al., 2016; Bouazizi and Ohtsuki, 2016; Shah et al., 2016b; Wang et al., 2016; Waters et al., 2019]. Em relação às técnicas não-supervisionadas, são utilizadas quando não é possível ter um conjunto de dados rotulados para que o modelo treine e classifique as sentenças/opiniões. Geralmente cada autor tenta desenvolver o próprio algoritmo para a criação de um modelo mais eficiente como Sui et al. [2012].

Existem, também, técnicas híbridas que combinam técnicas não-supervisionadas e supervisionadas. Em alguns casos, essas técnicas apresentam uma acurácia maior em relação às técnicas puramente supervisionadas [Wang et al., 2016]. Isso porque, elas fazem usos de dicionários léxicos no *dataset* e buscam gerar mais características aos modelos de entrada dos classificadores.

No trabalho de Kouloumpis et al. [2011], é feita a análise de sentimento e identificação de polaridade para *tweets* relacionados a certos tópicos (*hashtags*). Foi demonstrado um modelo de aprendizagem de máquina supervisionada que classificava os *tweets* em positivo, negativo e neutro. Nesse trabalho foram utilizados três *datasets*, primeiro um *dataset* contendo apenas *hashtags* e sua classificação rotulados, um *dataset* de *emoticons* com suas respectivas classificações, e por fim um *dataset* de 4000 *tweets* manualmente rotulados.

O processo de classificação do modelo funciona como a maioria dos trabalhos com esse objetivo, há uma fase de pré-processamento onde é realizada a tokenização (Quebra dos termos do texto em pedaços), normalização, e a *part-of-speech* (POS tagging).

Para os experimentos, foram utilizadas algumas características, por ser o padrão, foram utilizados unigramas e bigramas, e também características comumente utilizadas em análise de sentimento para representar os dados lexicamente e as características de *part-of-speech* (POS).

O classificador utilizado foi o Adaboost.mh [Kégl, 2014] com 500 rodadas de *boost*. Esse processo foi rodado dez vezes, e foi tirada uma média de acurácia. Vale ressaltar que os autores também utilizaram o SVM, que obteve resultados parecidos, mas menos acurados no geral. Eles rodaram os experimentos utilizando n-gramas, n-gramas e características léxicas, n-gramas e características de POS, n-gramas e características do Twitter (abreviações comumente utilizadas no Twitter) [Wasden, 2006], e por fim todas juntas. Os experimentos que obtiveram melhores resultados foi a combinação dos três conjuntos de características (n-gramas + léxicos + twitter) com uma média de acurácia de 75%.

Shah et al. [2016a] demonstram um modelo de classificação de *tweets* em positivos, negativos e neutros. A arquitetura desse modelo é semelhante aos trabalhos utilizados como *baseline* nessa área, porém ele adiciona alguns tratamentos a mais que visam melhorar a acurácia, são eles a classificação de *hashtags* e análise de *emoticons*. Abaixo a arquitetura resumida do modelo:

Nesse trabalho, a análise de *emoticon* é básica, sendo dado a cada *emoticon* básico (felicidade, tristeza, e surpresa) o rótulo correspondente (positivo e negativo). A acurácia máxima obtida por eles foi de 81%.

Bahrainian and Dengel [2013] apresentam uma ferramenta para sumarização de sentimentos e classificação de *tweets* em positivo, negativo e neutro. Por exemplo, dado a entidade alvo "iPhone"realizar uma análise desse tópico no Twitter. O modelo desse trabalho é híbrido, utilizando aprendizagem de máquina supervisionada para a tabulação de sentimentos (palavras que expressam tristeza, felicidade, raiva, surpresa, entre outros) e também a parte não-supervisionada para a detecção de polaridade (positivo e negativo) do *tweet*, essa é utilizada como uma característica adicional ao modelo supervisionado. Como oportunidades, poderíamos melhorar a acurácia da classificação dos *tweets*, e fazer isto de forma eficiente, e de maneira mais geral possível, pois modelos supervisionados funcionam bem quando a base é extremamente grande (*tweets* do ano todo) o que se torna inviável a coleta e a rotulação da mesma. Já a não-supervisionada tem uma acurácia menor, e um processamento maior, pois faz processamento léxico e sintático das sentenças a partir de um *target* (ou não).

Terrana et al. [2014] apresentam um método não-supervisionado para classificação de *tweets* em positivo, negativo e neutro. Eles fazem isso unicamente a partir de *emoticons* presentes nos *tweets*, evitando assim qualquer interferência humana durante

o processamento. Como contribuição, é apresentado o método deles, o algoritmo de execução do programa e mostra que é possível analisar os *tweets* em tempo real unicamente a partir dos *emoticons* presentes nos mesmos. Eles fazem a polarização do *tweet* em tempo de execução. O resultado deles é comparado com os demais modelos presentes na literatura, utilizando abordagem supervisionada e híbrida, todavia a acurácia desse método não é tão boa quanto algumas híbridas. A acurácia média para o modelo deles foi de 72%.

Kang et al. [2012] demonstram uma abordagem híbrida para a classificação de polaridade de avaliações de restaurantes. Foram utilizadas unigramas e bigramas para a representação das características. O desempenho dessa abordagem, foi em média 72%, sendo que ela se mostrou muito boa na classificação de avaliações negativas, mas muito ruim com relação às positivas.

Ye et al. [2009] apresentam uma abordagem para classificação de polaridade (negativo, positivo) referentes a avaliações de vôos e viagens em sites de turismo. Eles utilizam n-gramas para representação das características extraídas, e utilizam os algoritmos SVM e Naive Bayes para a classificação. Os resultados obtidos com relação à acurácia média ficaram na média de 80%.

Stein et al. [2019] demonstra uma abordagem hierárquica para a classificação de polaridade, usando como forma de representação de modelos/características conhecida como Words Embeddings, no trabalho deles, foi utilizada um vocabulário gerado pela biblioteca FastText. O modelo deles é híbrido, onde dependendo do nível da classificação são utilizados algoritmos supervisionados e não supervisionados. Em seu trabalho, foram testadas diversas combinações de algoritmos para os níveis, onde a melhor combinação obteve uma acurácia de 92%.

3.2 Comentários finais do capítulo

Os trabalhos atuais citados ao longo desse capítulo apresentam diversas abordagens na tarefa de análise de sentimentos e classificação de polaridade em ambientes online. No entanto, não foi definido nenhum método que seja capaz de classificar textos curtos apenas com *emojis* de maneira eficiente. Desta forma, torna-se uma oportunidade de estudo viável, observando as limitações da literatura. A Tabela 3.1 apresenta um resumo dos principais trabalhos relacionados.

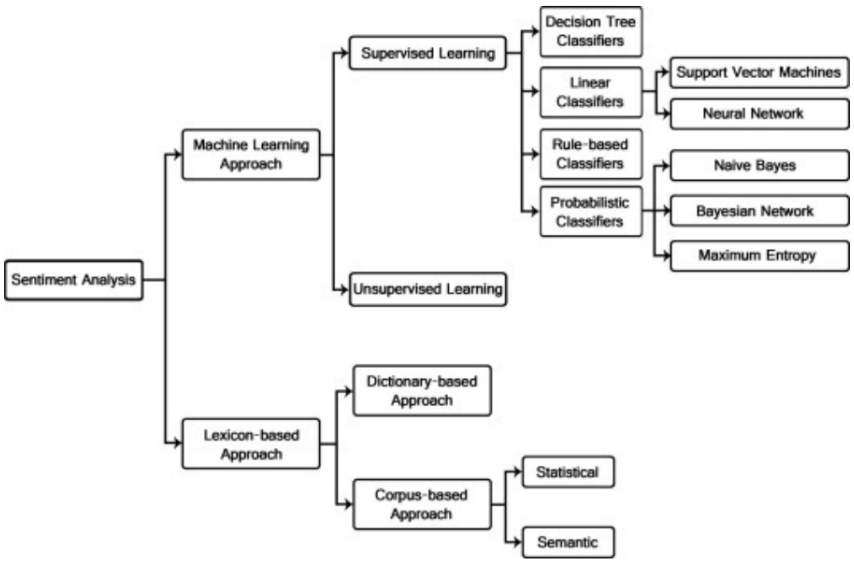


Figura 3.1: Técnicas de classificação de sentimento Medhat et al. [2014]; Serrano-Guerrero et al. [2015b].

Tabela 3.1: Resumo dos trabalhos relacionados

Autor	Rep. de Dados	Híbrido	Alg. de Classificação	Acurácia
Kouloumpis et al. [2011]	Unigramas e Bigramas	Sim	Adaboost.MH	83%
Shah et al. [2016a]	Unigramas e Bigramas	Não	Naive Bayes	81%
Terrana et al. [2014]	Frequência de Termos	Não	Estatístico usando apenas TF	60%
Bahrainian and Dengel [2013]	Term Frequency-Inverse Document Freq.	Sim	SVM	87%
Wang et al. [2005]	Termos e suas Probabilidades	Sim	Naive Bayes	69%
Wilson et al. [2005]	Termos	Não	Naive Bayes	65%
Sui et al. [2012]	Termos com seus pesos	Não	SVM	77%
Medhat et al. [2014]	Termos da sentença	Sim	Baseado em técnicas PLN	-
Kang et al. [2012]	Unigramas e Bigramas	Sim	Naive Bayes e SVM	72%
Ye et al. [2009]	N-gramas	Não	Naive Bayes e SVM	80%
Stein et al. [2019]	Word Embedding	Não	CNN (Redes neurais) e SVM	92%

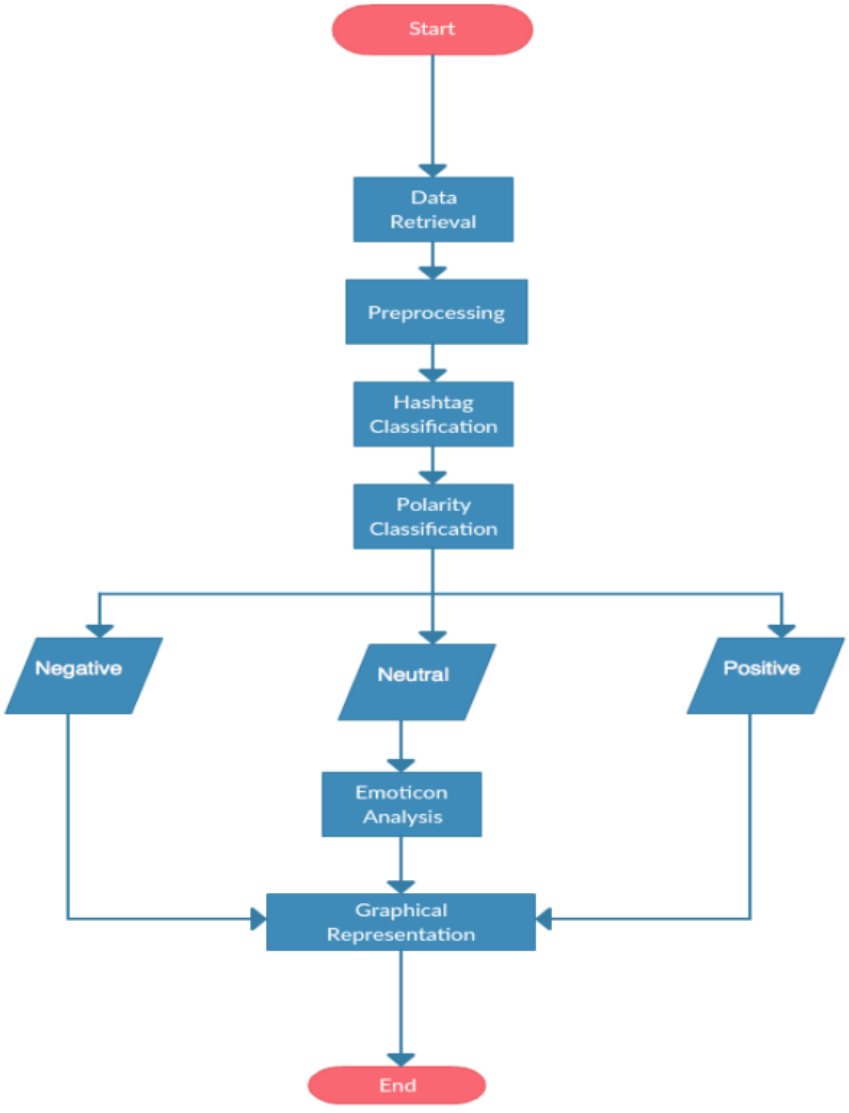


Figura 3.2: Arquitetura [Shah et al., 2016a].

Capítulo 4

Abordagem Proposta

De forma sucinta, a abordagem proposta, ilustrada na Figura 4.1, busca inferir a polaridade de avaliações a respeito de produtos e serviços publicados em ambientes virtuais, dando ênfase na utilização de emoticons e emojis e quantificando a riqueza linguística deste vocabulário para expressar polaridade e opinião. Para atingir esse objetivo a abordagem proposta inclui as seguintes etapas:

1. Seleção de categorias e dados;
2. Coleta e armazenamento dos dados;
3. Transcrição dos emojis/emoticons em palavras;
4. Adição de colunas para dados transcritos;
5. Pré-processamento;
6. Balanceamento;
7. Filtro de características TF-IDF com n-gramas;
8. Execução algoritmos de classificação usando k-fold;
9. Coleta de métricas de saída dos algoritmos para cada fold;
10. Criação de gráficos dos resultados e escrita em arquivos.

4.1 Seleção e Coleta de dados

Ambientes virtuais de lojas online de produtos e serviços (e.g. Google Play e Amazon) são plataformas que permitem que usuários compartilhem suas opiniões e experiências através de avaliações públicas. A coleta destes dados públicos é particularmente interessante para validar técnicas de Análise de Sentimentos, pois essas avaliações possuem descrições textuais das opiniões e notas que quantificam com valores discretos (e.g. uma a cinco estrelas) a satisfação do usuário relatada no texto. Portanto, as próprias notas podem ser utilizadas como rótulos mais fidedignos das opiniões (sob a ótica do usuário). Como o objetivo deste trabalho é quantificar a riqueza de *emojis* e *emoticons* para determinar a polaridade de avaliações, foram consideradas apenas avaliações cujos textos possuíam pelo menos um *emoji*/*emoticon*. O detalhamento da base de dados e sua construção é detalhado no capítulo 5.

4.2 Pré-Processamento

Após a coleta dos dados, os documentos (avaliações) são tratados para eliminar atributos não representativos das avaliações que foram coletadas. O pré-processamento proposto, ilustrado na Figura 4.2, inclui as seguintes fases:

- A detecção de idioma descarta textos que não estejam em inglês;
- A função *Tokenizer* separa os documentos em *tokens* (segmentos de sentenças), permitindo processar os segmentos individualmente;
- A remoção de Menções elimina as citações a outros usuários e *hashtags* presentes nos documentos para que sejam avaliados apenas o texto e emojis/emoticons;
- As URLs foram removidas pelo mesmo motivo;
- A remoção de *stop words* elimina termos não representativos do documento (e.g. artigos, preposições e números).

A seguir, as avaliações textuais são representadas usando a abordagem *bag-of-words* (BoW), onde cada documento é representado por um vetor de palavras que o compõe computando os valores de TF-IDF (*term frequency, inverted document frequency*) [Aisopos et al., 2012]. A abordagem geral de BoW com TF-IDF é a principal técnica utilizada em Análise de Sentimentos sobre textos e maiores detalhes podem ser conferidos no trabalho de Aisopos et al. [2012].

Para quantificar a riqueza linguística de emojis/emoticons são consideradas três análises:

1. **Palavras + emojis** - Uma representação de *bag-of-words* com ambos os termos;
2. **Palavras** - Uma representação de *bag-of-words* apenas com palavras;
3. **Emojis** - Uma representação de *bag-of-words* apenas com *Emojis*.

4.3 Classificação

Nessa etapa, utilizamos a representação dos dados criada pela BoW com TF-IDF e submetemos ao algoritmo de aprendizagem de máquina. Vale ressaltar, que uma caraterística importante de técnicas supervisionadas é que elas apresentam fundamentalmente o mesmo comportamento: dado um conjunto de dados de treinamento, composto por instâncias que são formadas por vários atributos previamente rotulados (polaridades) nas classes de interesse [Moraes et al., 2013], o modelo aprende características (base de treino) de cada classe e poderá ser usado para classificar outras amostras (base de teste).

Nessa fase, a priori, optou-se por utilizar os algoritmos do estado da arte, visto que os mesmos já apresentaram resultados validados na literatura. Os algoritmos utilizados foram o SVM, Naive Bayes, Max Entropy e Random Forest.

Inicialmente, os métodos que serão utilizados são Unigramas e N-Gramas. Unigramas é o método mais simples de extração de características e é definido como: olhar para uma palavra de cada vez em um texto e extrair características. Vale ressaltar que este pode ser estendido a um N-grama, a fim de explorar a ordenação das palavras. Ele pode ser usado em diferentes estados de texto, como caracteres, palavras ou frases.

Por sua vez, o N-grama é definido como tendo um conjunto de palavras sequenciais em um texto, por exemplo, se $N = 2$, significa olhar para um par de palavras sequenciais de cada vez, que é chamado de bigrama.

A seguir, uma imagem representativa do modelo de classificação e os experimentos realizados:

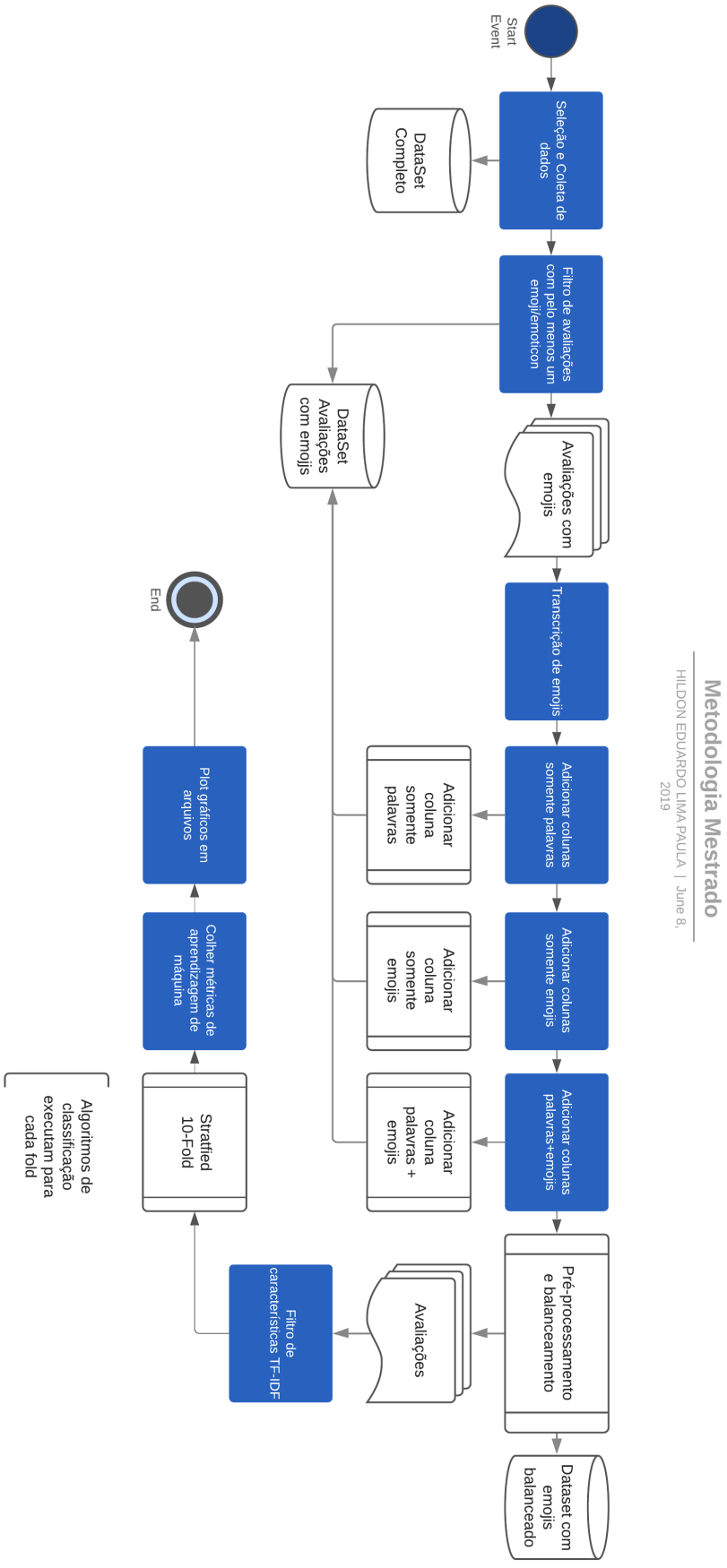


Figura 4.1: Abordagem proposta.

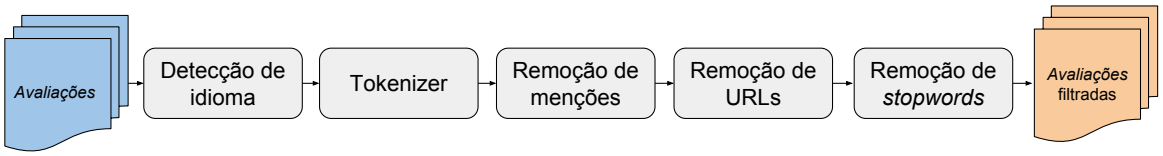


Figura 4.2: Etapas do pré-processamento.

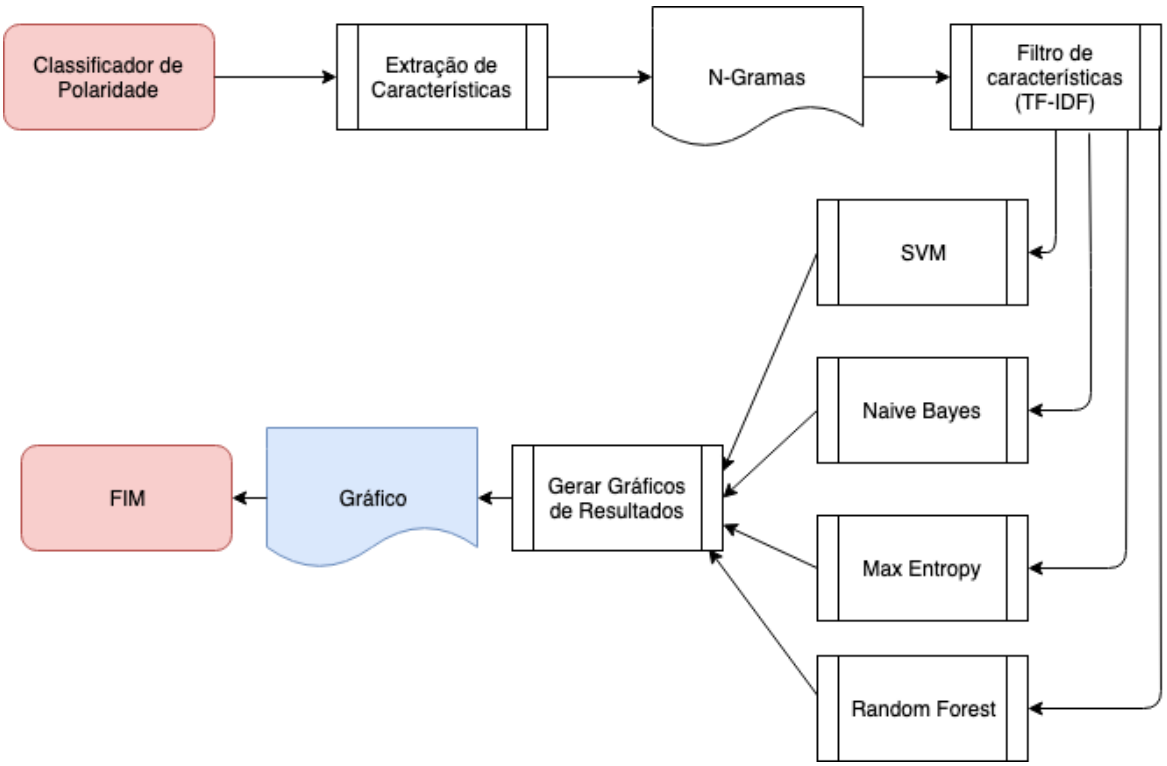


Figura 4.3: Metodologia e experimentos.

Capítulo 5

Metodologia

5.1 Estratégia de Pré-processamento

Como os *emojis* e *emoticons* não são palavras comuns (compostas de letras e acentos), os classificadores não conseguem extrair características de caracteres *unicodes*, e nem de símbolos e pontuações. Assim foi necessário a realização de diversos testes, visando uma maneira eficiente de mapear e representar os *emojis* e *emoticons*.

Inicialmente, os *emojis* haviam sido convertidos para strings de bytes, todavia, não obtivemos resultados satisfatórios. Uma vez que palavras com carácteres unicodes continham números e símbolos, o que atrapalhava os classificadores. Dessa forma, optamos por mapear os *emojis* em palavras diferentes, utilizando um código único para cada emoji/emoticon (sequencialmente segundo o alfabeto).

Na fase de pré-processamento os *emojis* foram então mapeados em palavras, conforme o apêndice A. Essa operação busca transcrever os *emojis*/*emoticons* para palavras mais fáceis de serem extraídas características.

Para cada tipo de pré-processamento, os experimentos foram realizados levando em consideração que as notas de 1 a 5 de cada avaliação eram sinais da polaridade da mesma, sendo assim, avaliações com notas 1 indicam que o usuário não ficou nada satisfeito (negativo), e nota 5 muito satisfeito (positivo). Parte dos trabalhos na área de classificação de polaridade classificam o texto em negativo, positivo e neutro [Almeida et al., 2016], os experimentos levaram apenas os extremos em si, tendo em vista que a maioria das plataformas de entretenimento que possuem avaliações estão migrando para este tipo de nota, como por exemplo o *Netflix* [Inc., 2017]. No caso do *Netflix*, a mudança no sistema de avaliação além de melhorar a acurácia geral do sistema, aumentou o engajamento dos usuários nas avaliações em 200 % [Prado, 2018]. Esses experimentos ajudaram a comparar o desempenho da abordagem, e observar onde ela

teve o melhor desempenho.

Os experimentos foram conduzidos utilizando técnicas de representação de características tradicionais, tendo como entrada as avaliações textuais e suas respectivas pontuações. A seguir, o texto é pré-processado e é executada a vetorização e transformação em uma BoW utilizando TF-IDF (*term frequency–inverse document frequency*) para identificação dos termos significativos de cada texto, e por fim este vetor de características é fornecido como entrada ao SVM.

5.2 Granularidade de Classes

A partir da base original, foram extraídas apenas as avaliações que continham *emojis* para avaliarmos se o uso do mesmo aumenta a chance de acerto na classificação. Avaliações com notas 1 e 2 receberam rótulo -1, enquanto as com nota 3, 4 e 5 receberam rótulo 1.

Este cenário considera duas possibilidades de nota para cada avaliação: -1 (negativo) e 1 (positivo). A classe da nota -1 contém 50% de avaliações oriundas da nota 1 e 50% da nota 2 da base original, sorteados aleatoriamente. De forma análoga, a classe da nota 1 contém 50% de avaliações oriundas da nota 4 e 50% da nota 5 da base original.

Essa divisão foi feita para tentar avaliar a abordagem proposta para diferentes aplicações. Naturalmente, problemas de classificação com muitas classes tendem a ser mais difíceis, uma vez que a possibilidade opções é maior, e dependendo da extração das características, algumas classes podem não ser bem discriminadas. Essa dificuldade de diferenciar bem as classes pode acontecer inclusive com um ser humano, durante a rotulação.

Desta forma o problema tem duas classes (positivo, e negativo), por isso, torna-se mais fácil a discriminação entre as classes. O uso de *emojis* neste cenário também tende a facilitar a classificação, uma vez que há uma grande variedade de *emojis/emoticons* que expressam bem os sentimentos positivos e negativos, discriminando bem os dois extremos entre a base. Na Figura 5.1 é possível observar a divisão de classes do problema e a quantidade total de avalições com emojis utilizadas.

5.3 Estratégia de Utilização de Emojis

Visando quantificar a relevância/contribuição de *emojis* e *emoticons* como recursos linguísticos (ainda que representem uma linguagem informal) para expressão de senti-

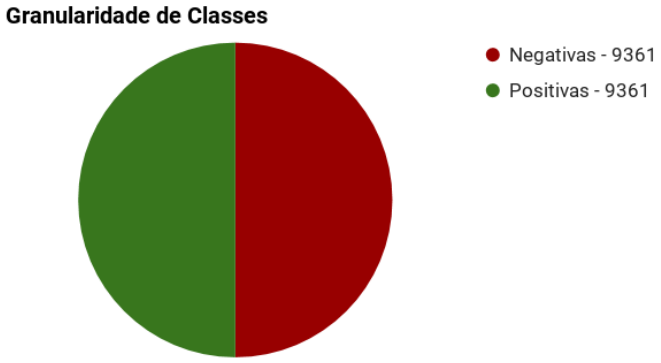


Figura 5.1: Polaridade das avaliações.

mentos e opiniões, cada cenário avalia a abordagem de Análise de Sentimentos considerando:

- **Somente Palavras**. Abordagem BoW tradicional que considera apenas palavras (*emojis* e *emoticons* são descartados).

- **Palavras+Emojis**. Abordagem que considera *emojis* e *emoticons* como palavras do vocabulário.

- **Somente Emojis**. Abordagem que considera somente os *emojis* e *emoticons* como palavras do vocabulário.

Para exemplificar as estratégias utilizadas para a quantificação da contribuição dos *emojis* e *emoticons*, temos a tabela 5.1.

Foram utilizadas duas formas de representação dos textos na BoW, unigramas e bigramas. N-grama é uma sequencia de N palavras, na sentença "I love this app <3<3 :D", podemos organizá-la conforme a tabela 5.2:

A partir de bigramas pra cima, podemos explorar as combinações dos pares de termos nas sentenças. Seria interessante avaliar a combinação de pares de emojis, contudo o tempo de processamento e extração de características tende a ser maior também.

Os experimentos adotam uma validação cruzada de dez grupos (*ten fold cross validation*). Os valores médios apresentados foram complementados com os intervalos de confiança para um grau de 95% de confiança. Esses intervalos são apresentados nos gráficos através de barras de erros.

5.4 Comentários finais do capítulo

Nesse capítulo foi apresentada a metodologia utilizada na realização dos experimentos, além de detalhar a construção da base a partir de avaliações de um ambiente online (Google PlayStore). Foi demonstrada também a divisão da base para os experimentos de polaridade (duas classes). Para este cenário foram executadas três abordagens de pré-processamento para o uso de *emojis*/*emoticons* na identificação de polaridade. Em todas elas, os *emojis* e *emoticons* foram convertidos para palavras enumeradas e definidas no apêndice A.

A primeira abordagem consistia em remover os *emojis*/*emoticons* das avaliações e executar o algoritmo. O segundo consistia em remover as palavras e manter apenas os *emojis*/*emoticons* e executar o algoritmo. O terceiro consistia em manter os dois na execução do algoritmo.

Este processo foi reproduzido com o uso de *bag-of-words* (BoW) em unigramas e bigramas. Não optamos por fazer trigramas e quadrigramas pois o bigrama se mostrou pior que o unigrama, nos levando a crer que para esse problema, quanto maior o valor n, pior o desempenho.

Tabela 5.1: Exemplo estratégia de uso de emojis e emoticons

Estratégia	Sentença original	Sentença processada
Palavras + Emojis	I love this app <3 :)	I love this app emoticonxbn emoticonxbx
Somente Palavras	I love this app <3 :)	I love this app
Somente Emojis	I love this app <3 :)	emoticonxbn emoticonxbx

Tabela 5.2: Exemplo de organização em unigramas e bigramas

Formato	Sentença
-	I love this app <3 :D
Unigrama	I, love, this, app, <3, :D
Bigrama	I love, love this, this app, app <3, <3 :D

Capítulo 6

Conjunto de Dados

A base de dados foi construída com avaliações coletadas da Google Play (loja de aplicativos móveis para Android). As avaliações foram coletadas utilizando a Google API[1], que permite captura de páginas de avaliações a partir do identificador do aplicativo e do idioma desejado. A Google API limita o acesso a 100 páginas de cada aplicativo por hora.

6.1 Análise da base sem discriminar categorias

Na Figura 6.1 é apresentada a distribuição das avaliações por categoria. Todas as avaliações possuem uma nota entre um e cinco. Ao fim da coleta de dados, foi obtido um total de 1.160.594 avaliações da Google Play, num total de 307 aplicativos distribuídos em 7 categorias: Entretenimento, Jogos, Mapa & Navegação, Vendas, Redes Sociais e apps de Clima.

A base de dados é composta de 1.160.595 linhas (contando o cabeçalho) e 10 colunas, descritas na Tabela 6.1.

Alguns dos campos da base de dados estão marcados na imagem 6.2. Todas as informações da base foram baixadas diretamente do site da Google Play, através de um crawler, o campo "category" também foi coletado de cada página do aplicativo.

Com a comparação na Figura 6.3 é possível observar o uso de emojis nas avaliações da base de dados.

O número de avaliações com emojis é 58.748 distribuídos entre as 7 categorias conforme a Figura 6.4.

[1] `https://github.com/facundoolano/google-play-scraper`.

Dentre as avaliações que continham ao menos um emoji, é possível observar quais são os emojis mais utilizados para cada estrela, conforme as Imagens: 6.5a, 6.5b, 6.6a, 6.6b, 6.7.

Após realizar um cálculo de distribuição cumulativa de frequência de emojis, podemos notar que a diferença dos emojis mais utilizados para os menos é bem alta, conforme podemos observar na Figura 6.8. Se a curva progride ao valor de Y igual 1 mais rápido, indica que os emojis mais utilizados estão muito mais distantes dos menos utilizados, sendo assim, se apenas os primeiros forem utilizados para a classificação a discriminação de classes será feita assertivamente e com menos emojis.

6.2 Análise da base por principais categorias

6.2.1 Jogos

A base de dados contém avaliações de 172 aplicativos/jogos, valor que corresponde a 56 % das avaliações. Das avaliações com emojis na base, 63,3 % são de Jogos, conforme observado na Figura 6.1. Ainda sobre as avaliações de jogos com emojis, é possível observar quais são os emojis mais utilizados para cada estrela, conforme as Imagens: 6.9, 6.10, 6.11, 6.12, 6.13.

Vale notar que a frequência de emojis em avaliações de jogos é a mais alta dentre as categorias, e também a categoria que mais contêm avaliações com emojis, algo que pode ser relacionado à faixa etária dos usuários que fazem as avaliações, uma vez que geralmente pessoas mais jovens tendem a consumir muito este tipo de aplicativo e da mesma forma, jovens tendem a usar mais emojis/emoticons [Marketing, 2019].

Após realizar um cálculo de distribuição cumulativa de frequência de emojis, podemos notar que a diferença dos emojis mais utilizados para os menos é bem alta, conforme podemos observar na Figura 6.14. A partir deste gráfico, pode-se inferir que quando as curvas crescem muito rapidamente para 1, significa que um conjunto menor de emojis é muito utilizado naquela nota, ou seja, pode-se definir a nota com um conjunto bem menor de emojis, dando mais importância a eles e desconsiderando os que tem menor frequência, a fim de facilitar o processo de extração de características e consequentemente o desempenho dos classificadores.

6.2.2 Entretenimento

Ao analisar os gráficos de frequência, percebe-se que os emojis mais usados são também o de maior utilização com relação a toda base, a única diferença está na quantidade de

uso, uma vez que não há tantas avaliações de entretenimento.

Após realizar um cálculo de distribuição cumulativa de frequência de emojis, podemos notar que a diferença dos emojis mais utilizados para os menos é bem alta, conforme pode ser observado na Figura 6.20.

6.2.3 Produtividade

Na base há avaliações de 64 aplicativos de produtividade, correspondente a 20,8 % das avaliações a segunda maior categoria da base. Por ser a segunda maior categoria em número de avaliações, a porcentagem de avaliações da mesma com emojis também é a segunda, onde 18,1 % de avaliações contém emojis. O rank de frequência de emojis por nota pode ser observado nas Figuras: 6.21, 6.22, 6.23, 6.24, 6.25.

Após realizar um cálculo de distribuição cumulativa de frequência de emojis, podemos notar que a diferença dos emojis mais utilizados para os menos é bem alta, conforme podemos observar na Figura 6.26.

6.2.4 Redes Sociais

Existem na base de dados avaliações de 26 aplicativos de Redes Sociais. É possível observar os emojis mais frequentes utilizados por nota nas Imagens: 6.27, 6.28, 6.29, 6.30, 6.31.

Após realizar um cálculo de distribuição cumulativa de frequência de emojis, podemos notar que a diferença dos emojis mais utilizados para os menos é bem alta, conforme podemos observar na Figura 6.32.

6.3 Comentários finais do capítulo

Neste capítulo foi apresentada a base de dados coletada para os experimentos, e uma análise dos dados nela contido. Na base de dados há uma predominância maior de avaliações de jogos (60%), isto acontece pois o foco maior foi coletar avaliações contendo emojis/emoticons, para que se possa avaliar qual a contribuição que estes elementos dão aos classificadores para o problema de classificação de polaridade em textos.

A partir do gráfico de distribuição cumulativa de frequência, pode-se observar que todas as categorias apresentam comportamento semelhante. Todas as curvas de notas de cada categoria tendem muito rapidamente para 1, ou seja, um conjunto menor de emojis são muito utilizados. Com isso, pode-se inferir que a partir de um conjunto menor de emojis é possível definir ou ajudar a definir melhor as notas, e os emojis

restantes, apesar de estarem presentes, podem ser desconsiderados ou ter um peso menor para os classificadores.

A base de dados está disponível on-line no link `https://drive.google.com/open?id=1l4Sf5wtZlyi9rZAn_JKGFQ99656g8Obl`, onde pode ser baixada e utilizada para fins acadêmicos e de pesquisas.

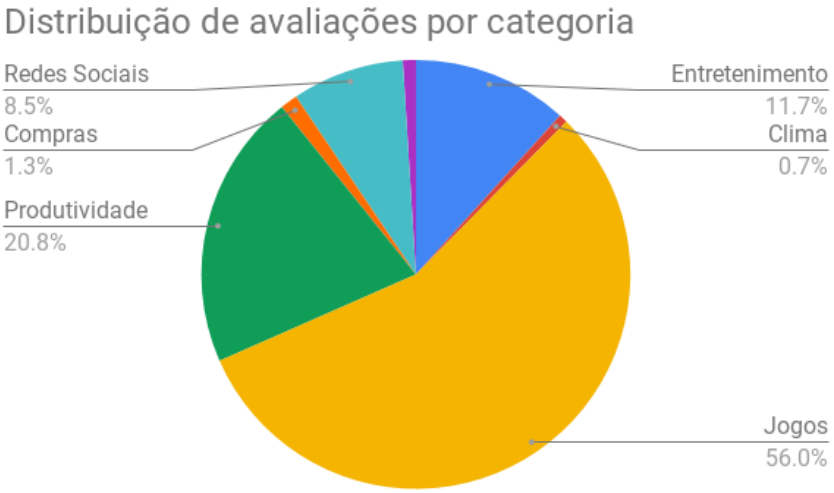
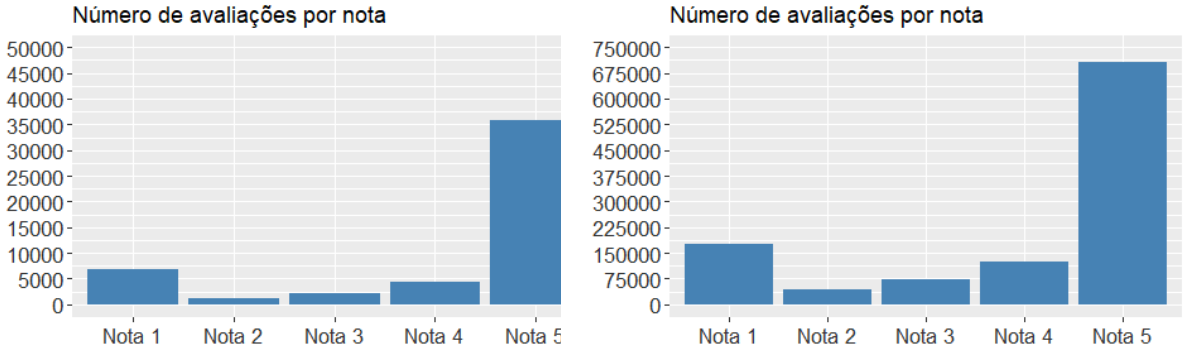


Figura 6.1: Distribuição de avaliações por categoria.

Coluna	Descrição
ID	Hash de identificação da avaliação
userName	Nome do usuário que escreveu a avaliação
img	1 se o usuário tiver imagem no perfil, 0 caso contrário
date	Data da avaliação no formato "MMM dd, yyyy"
score	Número de estrelas do review, valor inteiro entre 1 e 5
title	Título da avaliação
text	Conteúdo da avaliação
text_words_only	Texto da avaliação somente dom palavras (sem emojis)
text_emojis_only	Texto da avaliação somente com emojis (sem palavras)
text_emojis_and_words	Texto da avaliação com emojis e palavras
category	Categoria do aplicativo que foi avaliado

Tabela 6.1: Descrição de colunas do dataset

Figura 6.2: Exemplo de avaliação da Google Play.



(a) Número de avaliações com emojis. (b) Número de avaliações com e sem emojis.

Figura 6.3: Distribuição de dados para os dois cenários considerados.

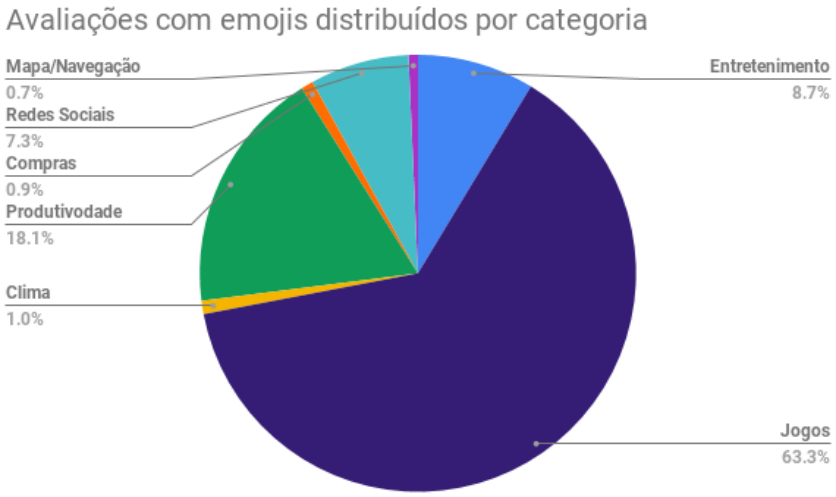
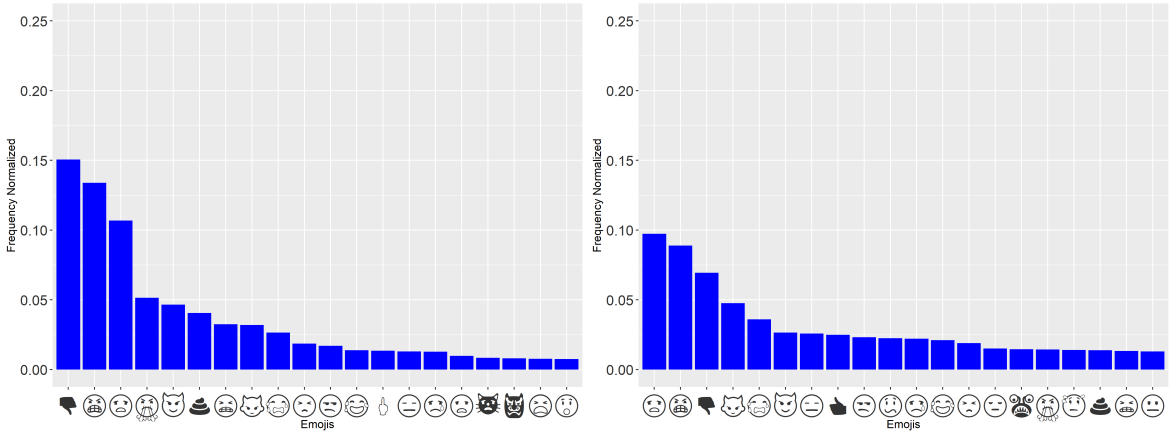
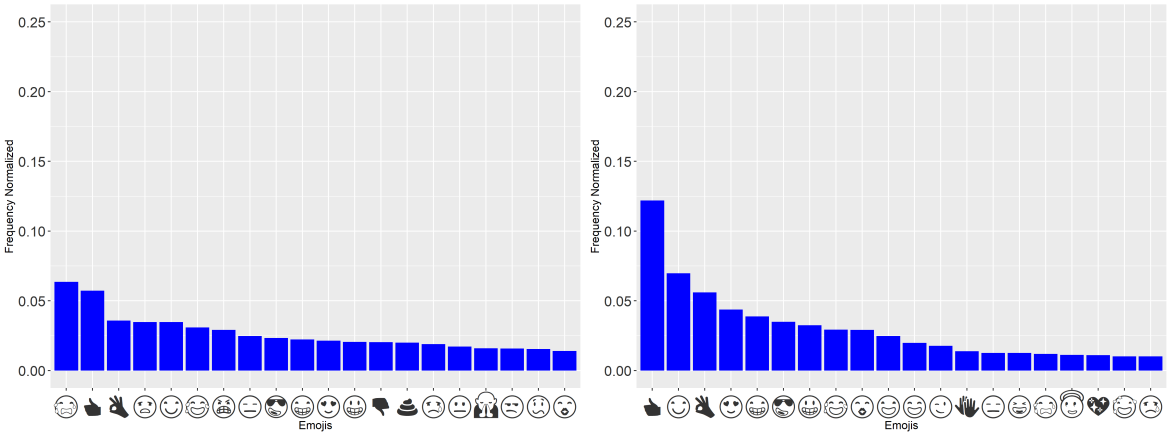


Figura 6.4: Distribuição de avaliações com emojis entre as categorias.



(a) Rank emojis mais utilizados com nota 1. (b) Rank emojis mais utilizados com nota 2.

Figura 6.5: Ranks emojis mais utilizados com nota 1 e 2.



(a) Rank emojis mais utilizados com nota 3. (b) Rank emojis mais utilizados com nota 4.

Figura 6.6: Ranks emojis mais utilizados com nota 3 e 4.

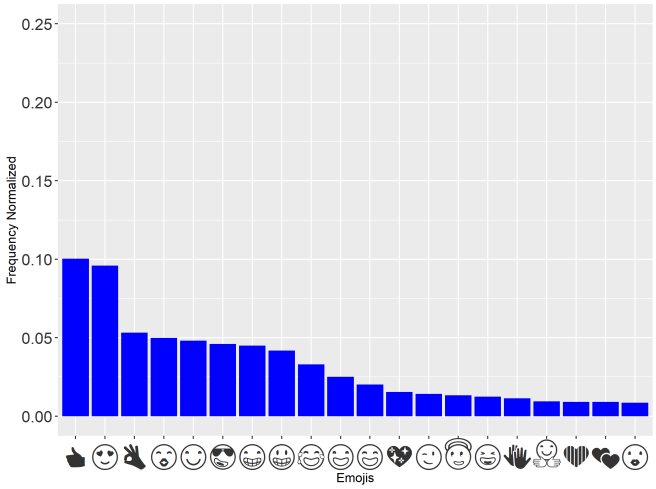


Figura 6.7: Rank emojis mais utilizados com nota 5.

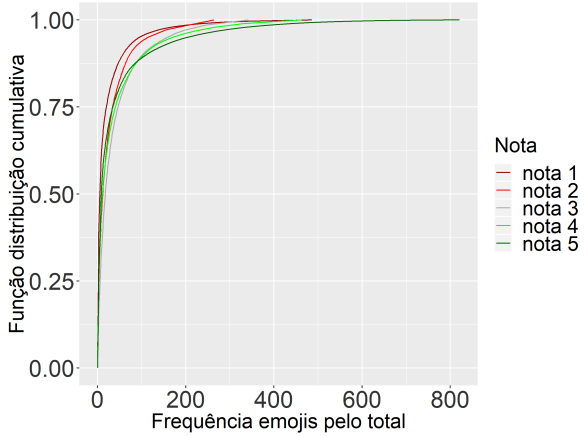
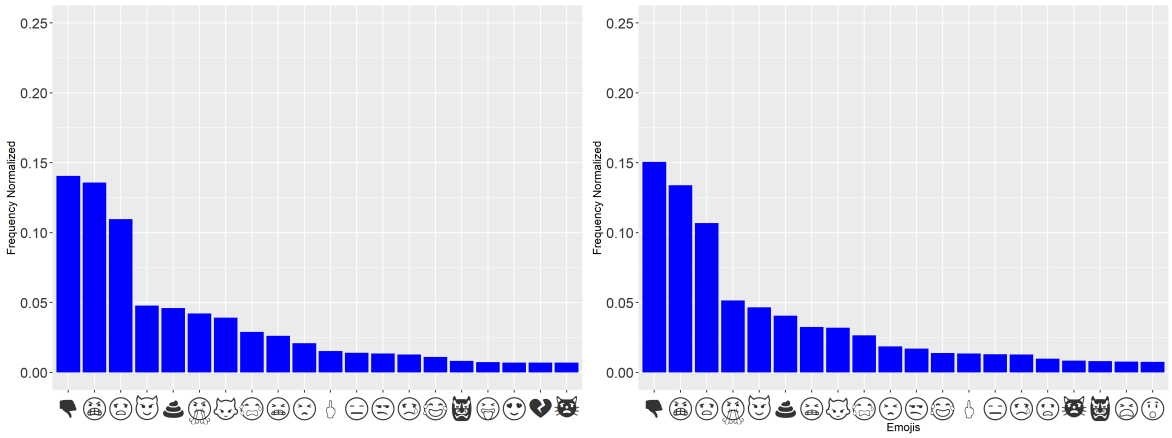


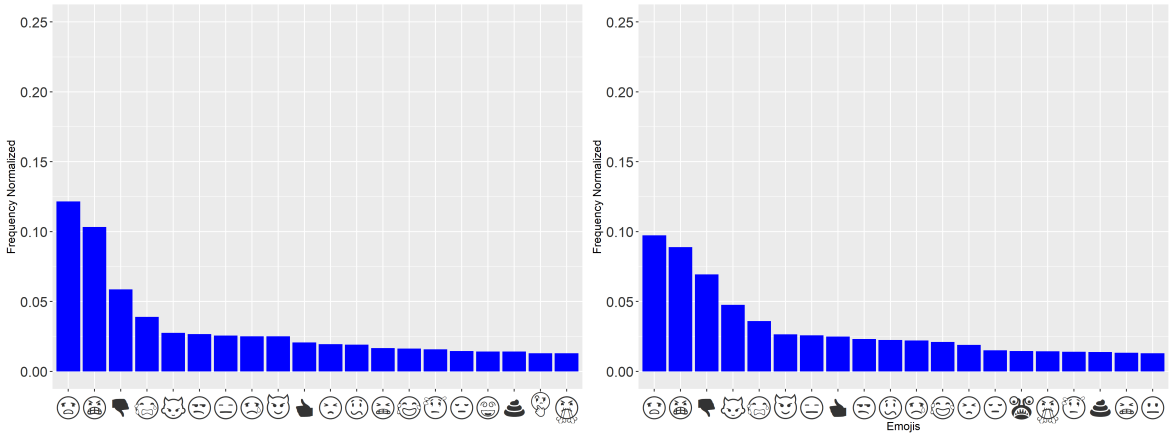
Figura 6.8: Distribuição cumulativa de frequência.



(a) Rank emojis mais utilizados de jogos com nota 1.

(b) Rank emojis mais utilizados de toda a base com nota 1.

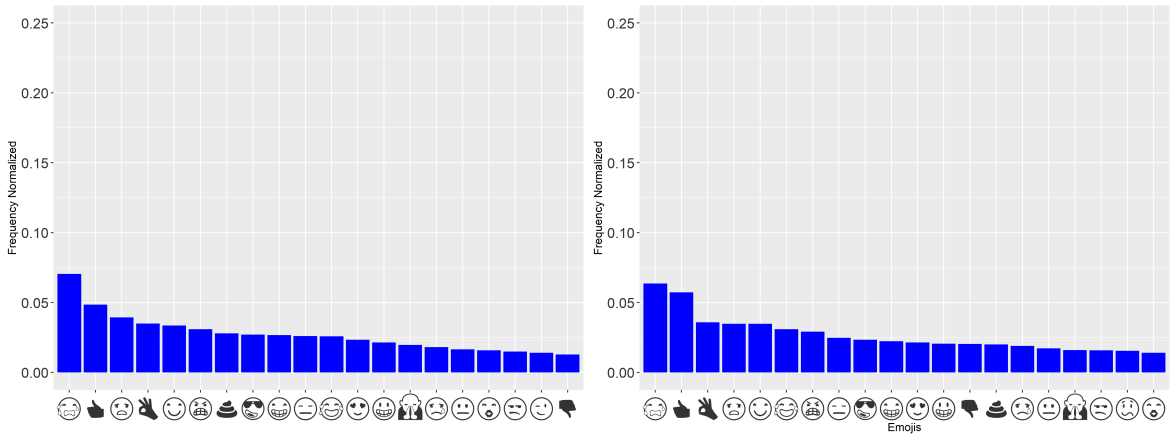
Figura 6.9: Comparativo de emojis mais utilizados em avaliações com nota 1 da categoria de jogos em relação à base toda.



(a) Rank emojis mais utilizados de jogos com nota 2.

(b) Rank emojis mais utilizados de toda a base com nota 2.

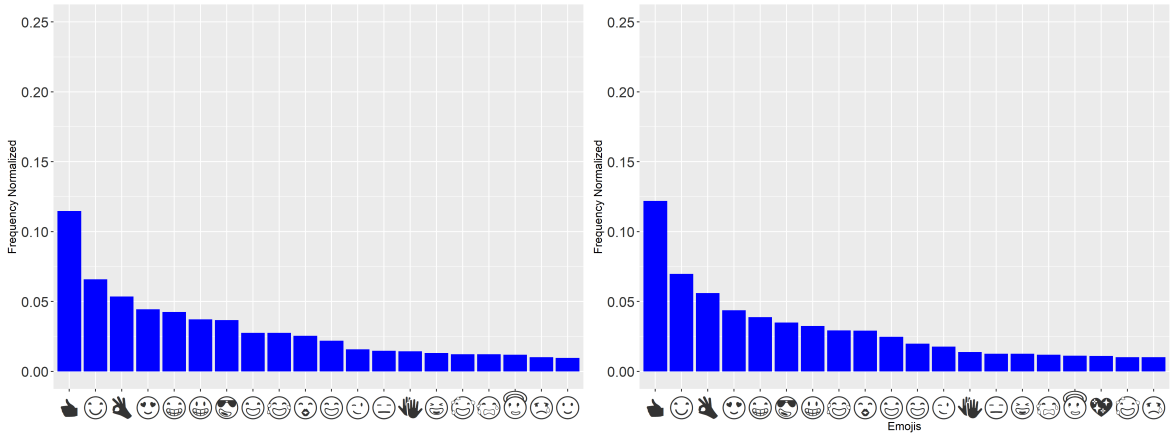
Figura 6.10: Comparativo de emojis mais utilizados em avaliações com nota 2 da categoria de jogos em relação à base toda.



(a) Rank emojis mais utilizados de jogos com nota 3.

(b) Rank emojis mais utilizados de toda a base com nota 3.

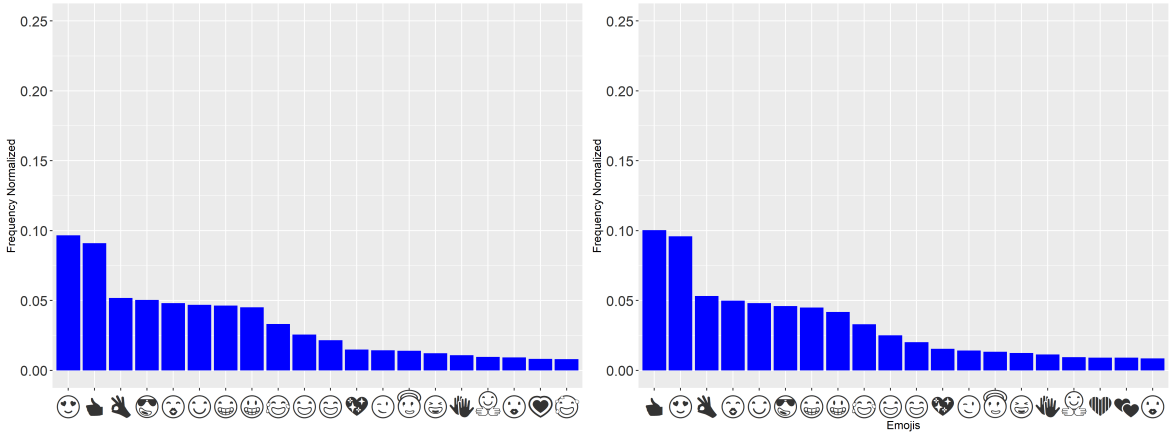
Figura 6.11: Comparativo de emojis mais utilizados em avaliações com nota 3 da categoria de jogos em relação à base toda.



(a) Rank emojis mais utilizados de jogos com nota 4.

(b) Rank emojis mais utilizados de toda a base com nota 4.

Figura 6.12: Comparativo de emojis mais utilizados em avaliações com nota 4 da categoria de jogos em relação à base toda.



(a) Rank emojis mais utilizados de jogos com nota 5.

(b) Rank emojis mais utilizados de toda a base com nota 5.

Figura 6.13: Comparativo de emojis mais utilizados em avaliações com nota 5 da categoria de jogos em relação à base toda.

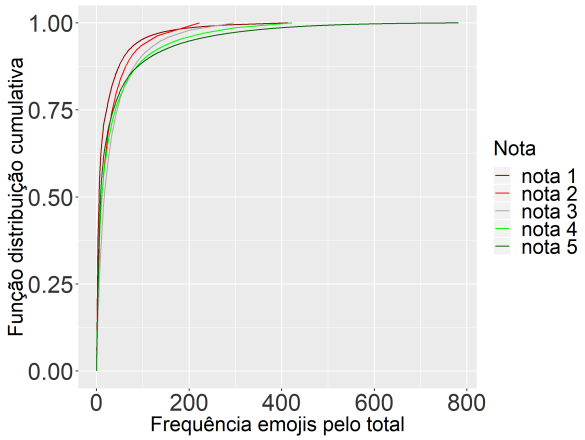
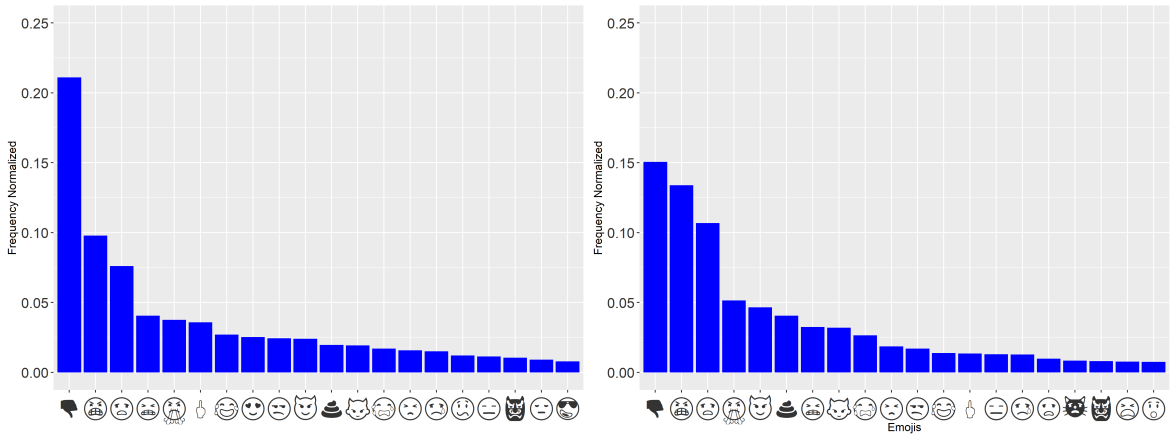


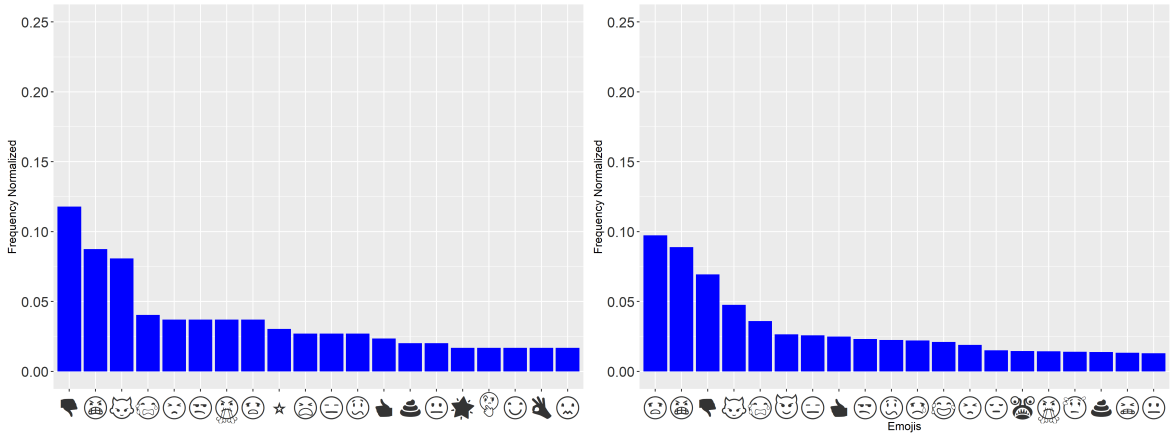
Figura 6.14: Distribuição cumulativa de frequência.



(a) Rank emojis mais utilizados de entretenimento com nota 1.

(b) Rank emojis mais utilizados de toda a base com nota 1.

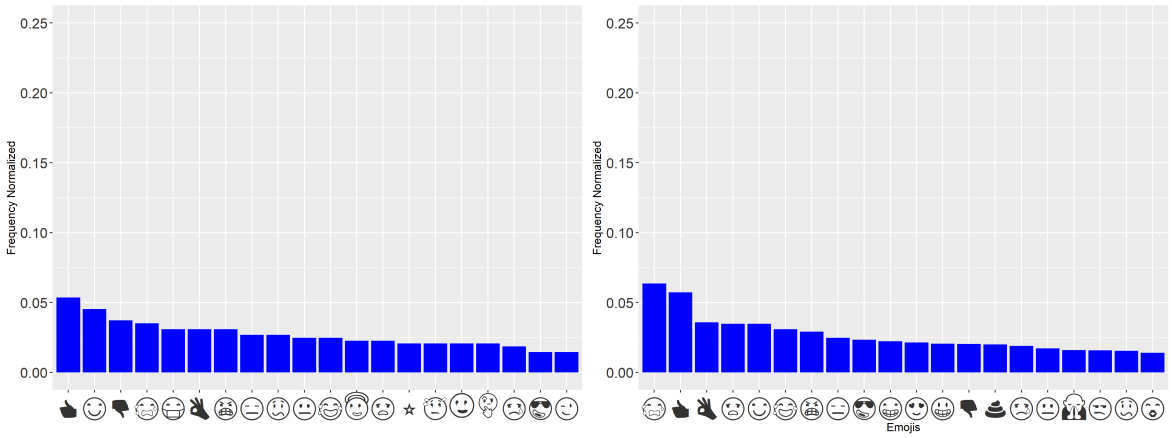
Figura 6.15: Comparativo de emojis mais utilizados em avaliações com nota 1 da categoria de entretenimento em relação à base toda.



(a) Rank emojis mais utilizados de entretenimento com nota 2.

(b) Rank emojis mais utilizados de toda a base com nota 2.

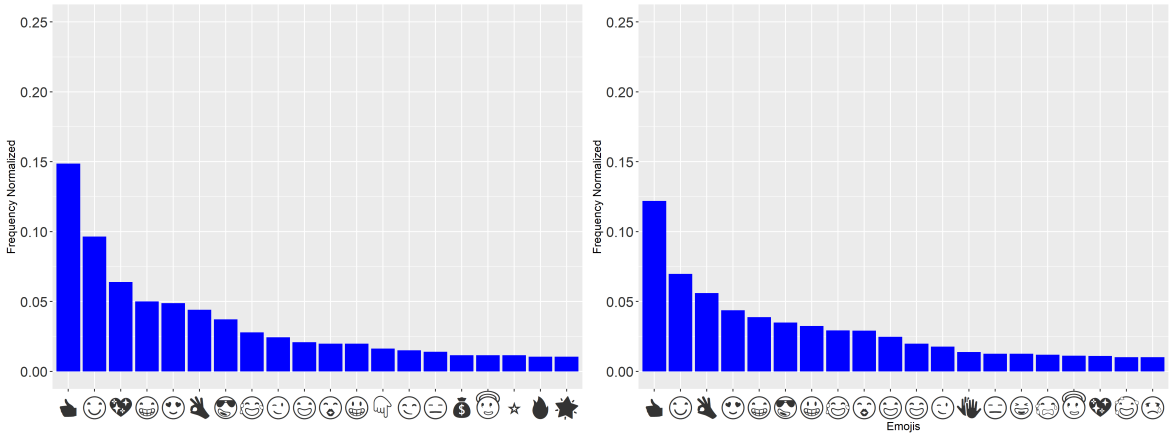
Figura 6.16: Comparativo de emojis mais utilizados em avaliações com nota 2 da categoria de entretenimento em relação à base toda.



(a) Rank emojis mais utilizados de entretenimento com nota 3.

(b) Rank emojis mais utilizados de toda a base com nota 3.

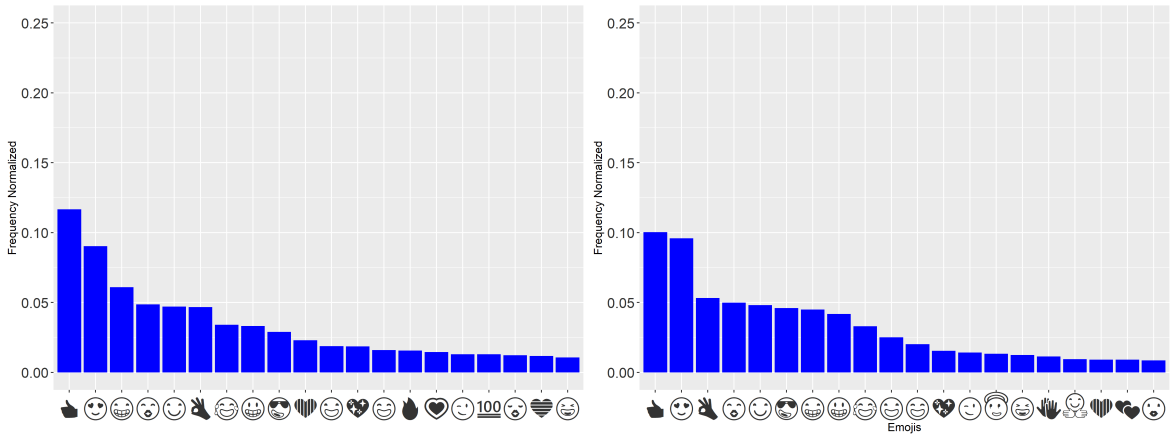
Figura 6.17: Comparativo de emojis mais utilizados em avaliações com nota 3 da categoria de entretenimento em relação à base toda.



(a) Rank emojis mais utilizados de entretenimento com nota 4.

(b) Rank emojis mais utilizados de toda a base com nota 4.

Figura 6.18: Comparativo de emojis mais utilizados em avaliações com nota 4 da categoria de entretenimento em relação à base toda.



(a) Rank emojis mais utilizados de entretenimento com nota 5.

(b) Rank emojis mais utilizados de toda a base com nota 5.

Figura 6.19: Comparativo de emojis mais utilizados em avaliações com nota 5 da categoria de entretenimento em relação à base toda.

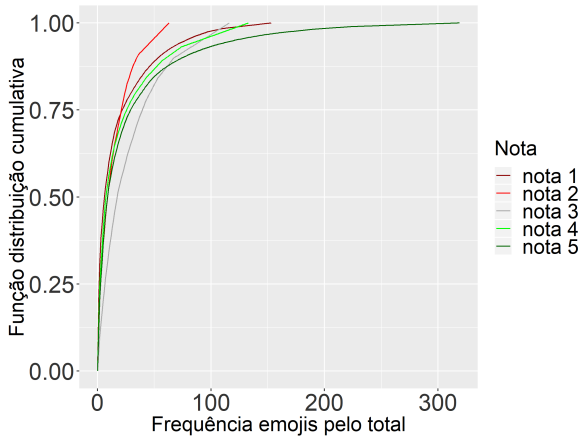
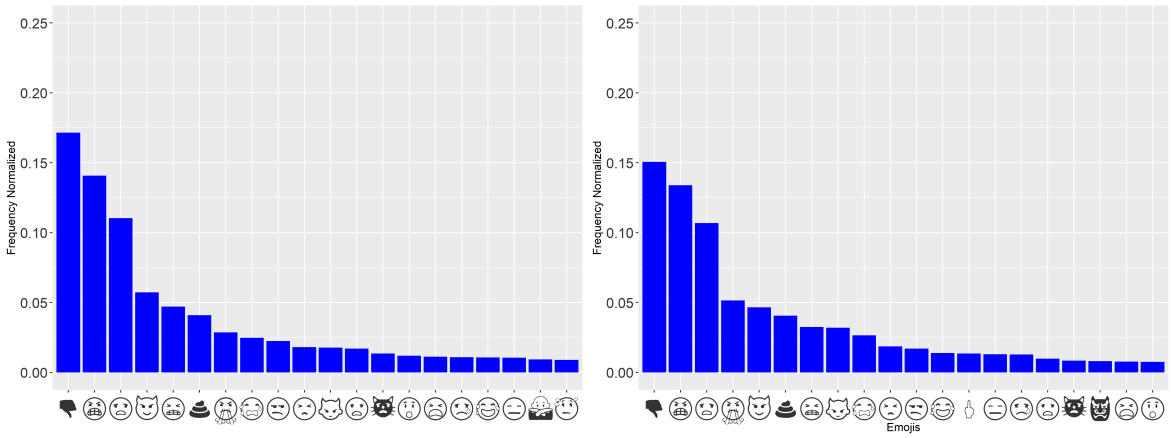


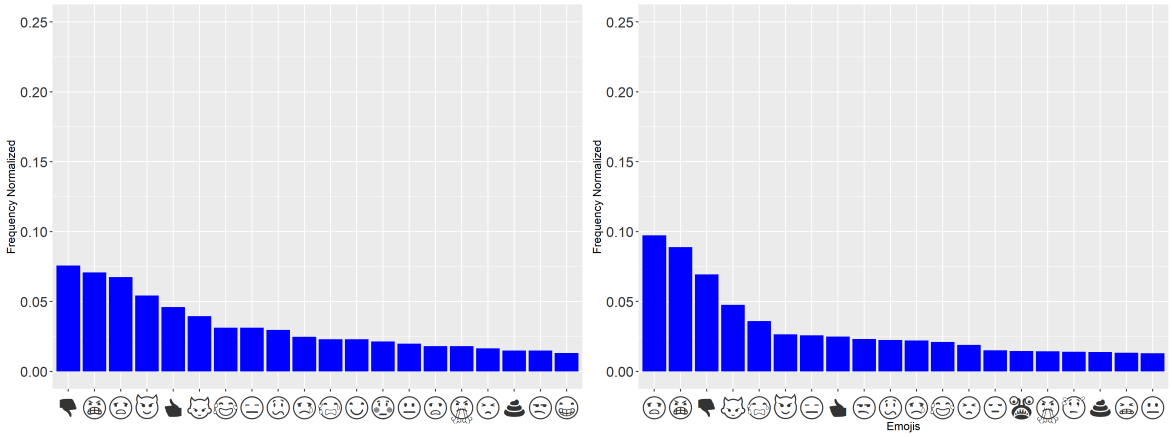
Figura 6.20: Distribuição cumulativa de frequência.



(a) Rank emojis mais utilizados de produtividade com nota 1.

(b) Rank emojis mais utilizados de toda a base com nota 1.

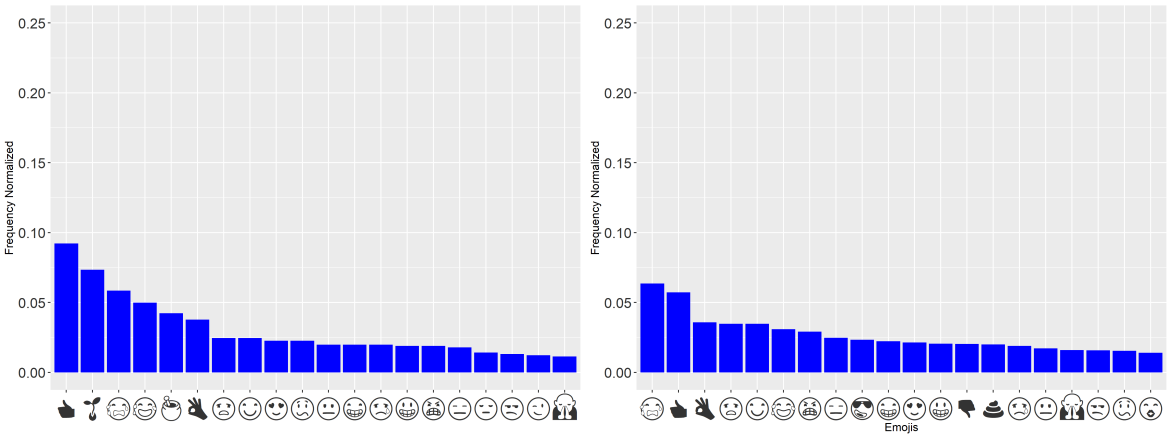
Figura 6.21: Comparativo de emojis mais utilizados em avaliações com nota 1 da categoria de produtividade em relação à base toda.



(a) Rank emojis mais utilizados de produtividade com nota 2.

(b) Rank emojis mais utilizados de toda a base com nota 2.

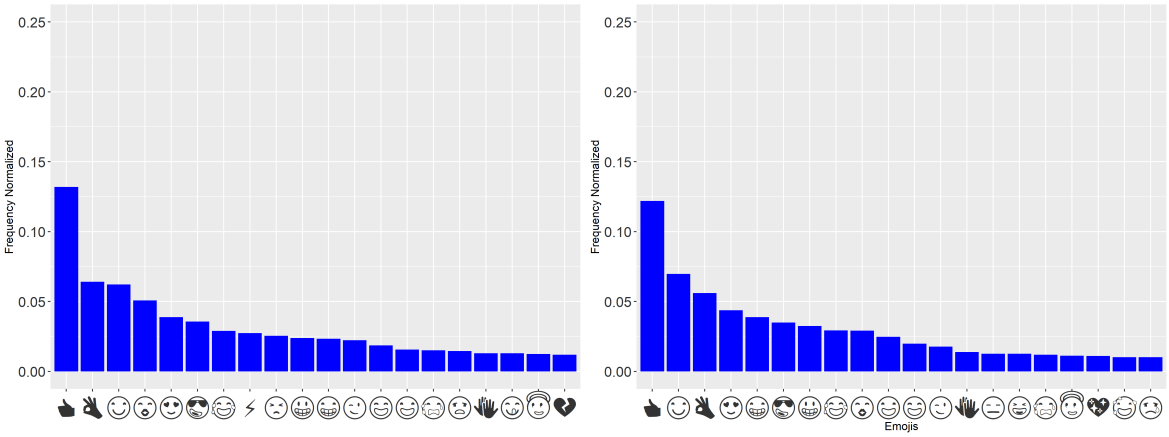
Figura 6.22: Comparativo de emojis mais utilizados em avaliações com nota 2 da categoria de produtividade em relação à base toda.



(a) Rank emojis mais utilizados de produtividade com nota 3.

(b) Rank emojis mais utilizados de toda a base com nota 3.

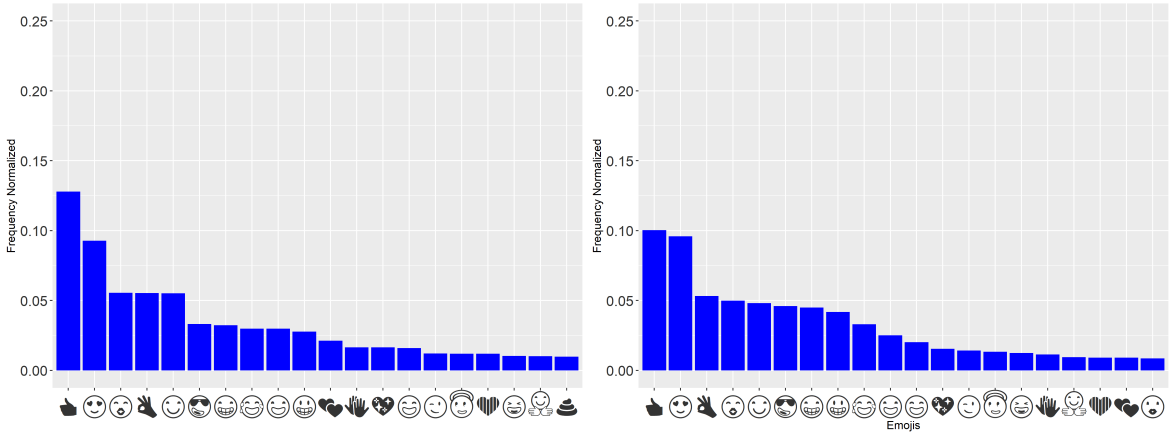
Figura 6.23: Comparativo de emojis mais utilizados em avaliações com nota 3 da categoria de produtividade em relação à base toda.



(a) Rank emojis mais utilizados de produtividade com nota 4.

(b) Rank emojis mais utilizados de toda a base com nota 4.

Figura 6.24: Comparativo de emojis mais utilizados em avaliações com nota 4 da categoria de produtividade em relação à base toda.



(a) Rank emojis mais utilizados de produtividade com nota 5.

(b) Rank emojis mais utilizados de toda a base com nota 5.

Figura 6.25: Comparativo de emojis mais utilizados em avaliações com nota 5 da categoria de produtividade em relação à base toda.

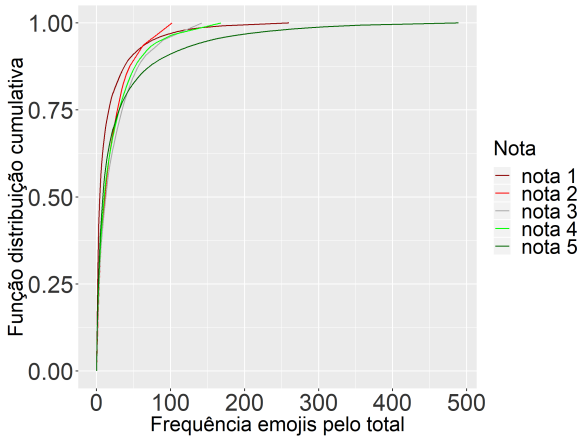
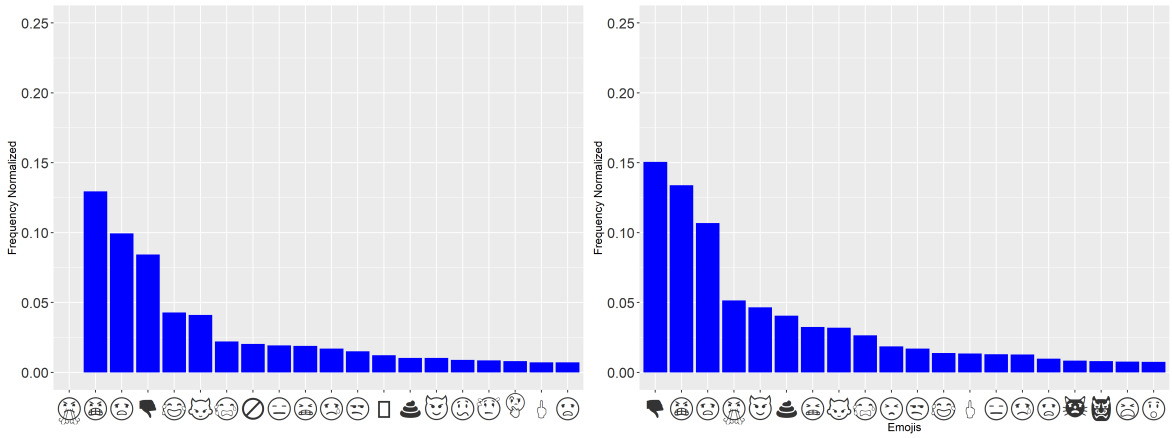


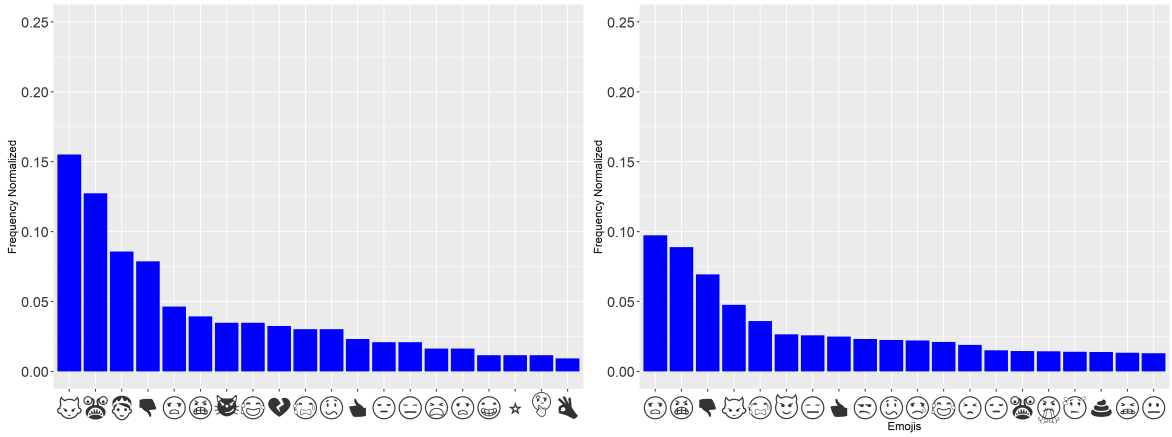
Figura 6.26: Distribuição cumulativa de frequência.



(a) Rank emojis mais utilizados de redes sociais com nota 1.

(b) Rank emojis mais utilizados de toda a base com nota 1.

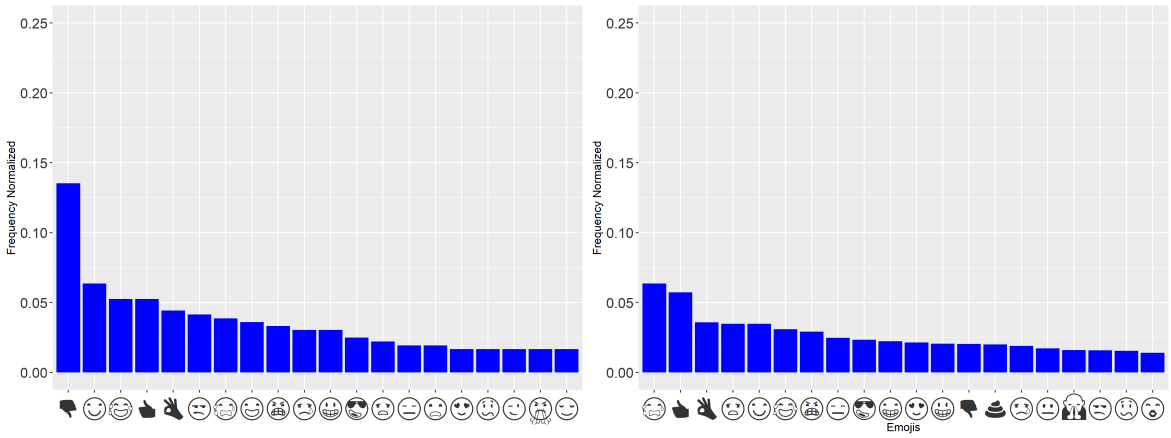
Figura 6.27: Comparativo de emojis mais utilizados em avaliações com nota 1 da categoria de redes sociais em relação à base toda.



(a) Rank emojis mais utilizados de redes sociais com nota 2.

(b) Rank emojis mais utilizados de toda a base com nota 2.

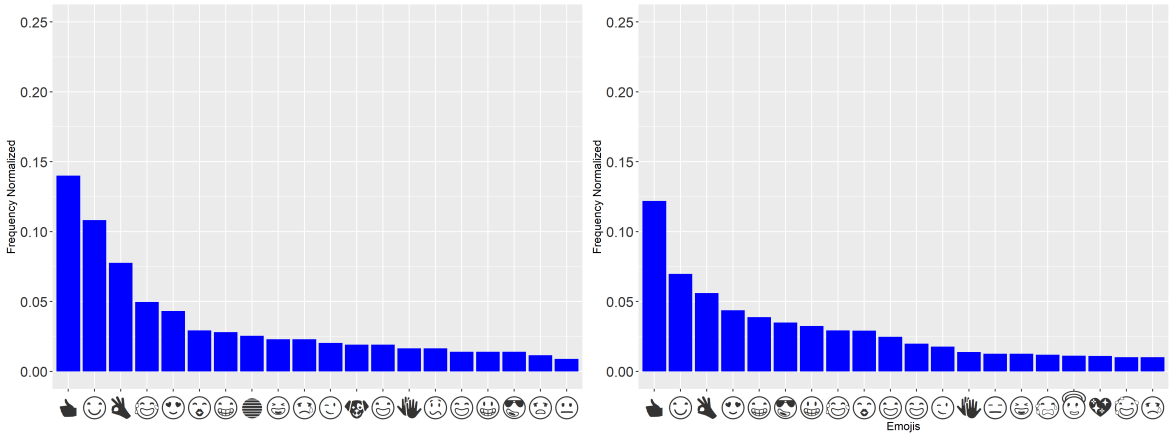
Figura 6.28: Comparativo de emojis mais utilizados em avaliações com nota 2 da categoria de redes sociais em relação à base toda.



(a) Rank emojis mais utilizados de redes sociais com nota 3.

(b) Rank emojis mais utilizados de toda a base com nota 3.

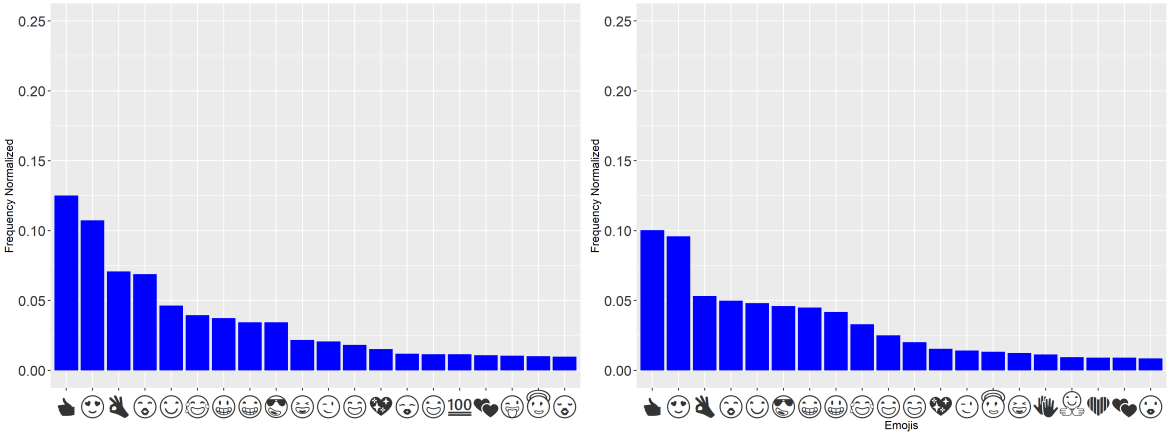
Figura 6.29: Comparativo de emojis mais utilizados em avaliações com nota 3 da categoria de redes sociais em relação à base toda.



(a) Rank emojis mais utilizados de redes sociais com nota 4.

(b) Rank emojis mais utilizados de toda a base com nota 4.

Figura 6.30: Comparativo de emojis mais utilizados em avaliações com nota 4 da categoria de redes sociais em relação à base toda.



(a) Rank emojis mais utilizados de redes sociais com nota 5.

(b) Rank emojis mais utilizados de toda a base com nota 5.

Figura 6.31: Comparativo de emojis mais utilizados em avaliações com nota 5 da categoria de redes sociais em relação à base toda.

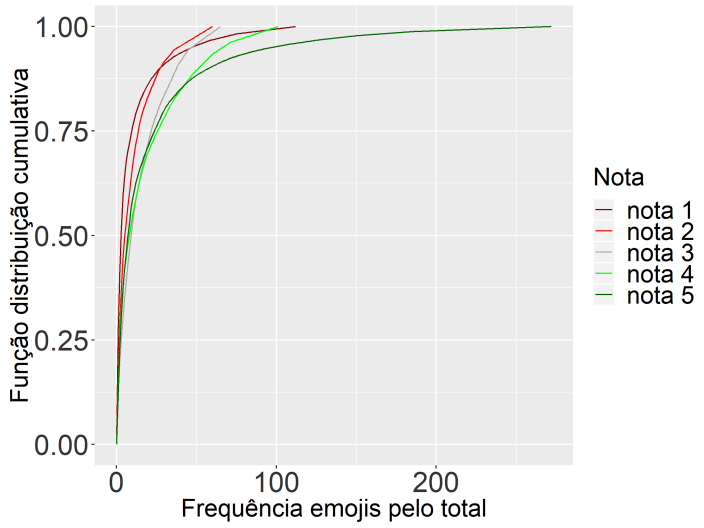


Figura 6.32: Distribuição cumulativa de frequência.

Capítulo 7

Experimentos e Resultados

Neste capítulo, são apresentados os experimentos utilizando os algoritmos SVM, Naive Bayes, Max Entropy e Random Forest. O capítulo esta distribuído pelo uso de Unigramas ou bigramas, e em cada um, dividido pela maneira de utilização dos emojis como entrada dos algoritmos de classificação: Somente emojis, somente palavras, e palavras e emojis juntos.

7.1 Unigramas

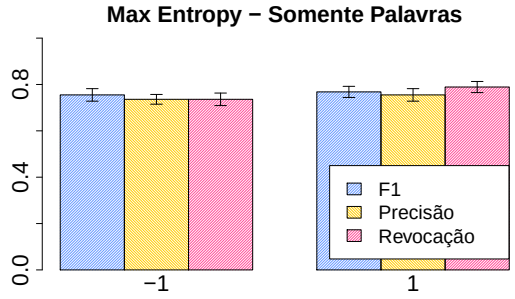
Nos experimentos mostrados a seguir as entradas para os algoritmos foram bag-of-words utilizando de unigramas, vetorizados utilizando o método TF-IDF. Ou seja, cada posição do vetor é composto por uma palavra e sua frequência/importância com relação a todas as avaliações.

7.1.1 Usando Apenas Palavras

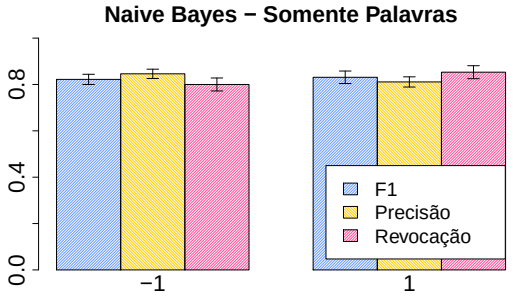
Nestes experimentos foram removidos todos os emojis/emoticons das avaliações, e submetidos ao modelo alternando os algoritmos, na Figura 7.1 são apresentados os gráficos com as métricas precisão, revocação e F1.

A Figura 7.2 apresenta as matrizes de confusão dos algoritmos. As matrizes mostram que ao utilizar somente palavras (abordagem tradicional), o algoritmo confunde a classe 1 (positivo) como -1 (negativo).

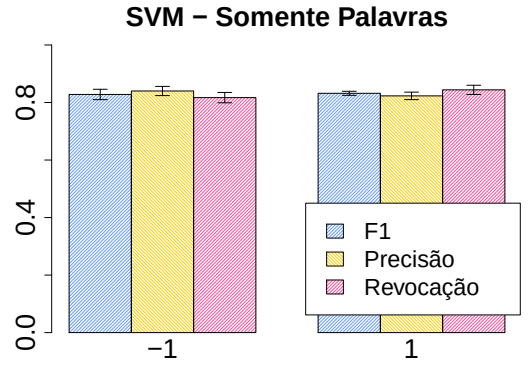
Os resultados confirmaram o estado da arte para este problema, com o uso de apenas palavras, o SVM obteve uma **acurácia geral média** foi **84,0% ± 1,6**, empatando tecnicamente com o Random Forest (83,0% ± 0,6) e o Naive Bayes (82,0% ± 0,2).



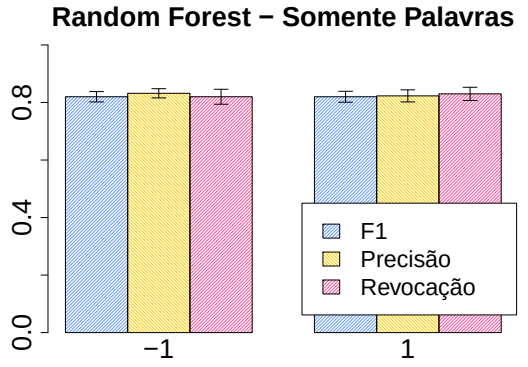
(a) Max Entropy.



(b) Naive Bayes.



(c) SVM.



(d) Random Forest.

Figura 7.1: Resultados dos Algoritmos ao usar somente palavras.

Na Figura 7.3 é feita a comparação entre os coeficientes de Mathew (MCC) resultado da execução de cada algoritmo [Matthews, 1975]. Ele leva em conta os verdadeiros e falsos positivos e negativos e é geralmente considerado como uma medida balanceada que pode ser usada mesmo se as classes forem de tamanhos muito diferentes. O MCC é, em essência, um valor de coeficiente de correlação entre -1 e $+1$. Um coeficiente mais próximo de 1 representa uma previsão perfeita, 0 uma previsão aleatória média e -1 uma previsão inversa.

7.1.2 Usando Apenas Emojis

Nos resultados a seguir, foram removidas as palavras do texto de cada avaliação e submetidos ao modelo, alternando apenas os algoritmos. Na Figura 7.4 são apresentados os gráficos com as métricas precisão, revocação e F1.

A Figura 7.5 apresenta as matrizes de confusão dos algoritmos. As matrizes mostram que ao utilizar somente emojis, os algoritmos obtem melhor resultado, a maioria dos algoritmos no entanto, não mantem uma eficácia de acerto por nota estável,

por exemplo, o Random Forest tem uma precisão de acerto de nota 1 de **90%** e de nota -1 **82% ± 0,1**.

Na Figura 7.6 é feita a comparação entre os coeficientes de Mathew (MCC) [Matthews, 1975] resultado da execução de cada algoritmo. Nestes experimentos, ambos os algoritmos tiveram resultados semelhantes para o coeficiente de Mathew, ou seja, todos eles tiveram uma assertividade alta ao prever os resultados, não chutando os resultados.

7.1.3 Usando Palavras + Emojis

Neste experimentos, foram combinadas palavras e os emojis transcritos. Na Figura 7.7 são apresentados os gráficos com as métricas precisão, revocação e F1.

A Figura 7.8 apresenta as matrizes de confusão dos algoritmos. As matrizes mostram que ao utilizar somente emojis, os algoritmos obtiveram os melhores resultados, a maioria dos algoritmos conseguiu manter uma estabilidade em relação à acurácia das avaliações de ambas as notas, sendo o SVM o melhor dentre eles.

Na Figura 7.9 é feita a comparação entre os coeficientes de Mathew resultado da execução de cada algoritmo. Ele leva em conta os verdadeiros e falsos positivos e negativos e é geralmente considerado como uma medida balanceada que pode ser usada mesmo se as classes forem de tamanhos muito diferentes.

7.1.4 Resumo dos resultados

Na Tabela 7.1 são apresentados os resumos dos resultados dos gráficos, separados por algoritmo e estratégia de uso dos emojis. O algoritmo SVM obteve o maior desempenho geral, onde a estratégia mais eficaz foi a combinação de palavras + emojis, atingindo a acurácia de **91,0% ± 0,1** para a métrica F1 de ambas as classes. Dentre as estratégias de uso, pode-se perceber que o uso de emojis facilita ainda mais o trabalho do classificador, dando um ganho de cerca de **3,0% ± 0,1** quando comparadas as estratégias somente palavras contra somente emojis, e quando comparado ao uso de palavras combinados com emojis, este ganho sobe para cerca de **9,0% ± 0,1**.

7.2 Bigramas

Nos experimentos mostrados a seguir as entradas para os algoritmos foram bag-of-words utilizando de bigramas, vetorizados utilizando o método TF-IDF. Ou seja, cada posição

do vetor é composto por duas palavras e sua frequência/importância com relação a todas as avaliações.

7.2.1 Usando Apenas Palavras

Nestes experimentos foram removidos todos os emojis/emoticons das avaliações, e submetidos ao modelo alternando os algoritmos, na Figura 7.10 são apresentados os gráficos com as métricas precisão, revocação e F1.

A Figura 7.11 apresenta as matrizes de confusão dos algoritmos. As matrizes mostram que ao utilizar somente emojis, os algoritmos obtiveram os melhores resultados, a maioria dos algoritmos conseguiu manter uma estabilidade em relação à acurácia das avaliações de ambas as notas, sendo o SVM o melhor dentre eles.

Na Figura 7.12 é feita a comparação entre os coeficientes de Mathew resultado da execução de cada algoritmo. Pelo gráfico, percebe-se que os algoritmos tiveram assertividade similar ao uso de unigramas para o mesmo cenário (apenas palavras), com exceção do algoritmo SVM, que com o uso de unigramas teve um aumento de $0,05$ em relação ao apresentado.

7.2.2 Usando Apenas Emojis

Nos resultados a seguir, foram removidas as palavras do texto de cada avaliação e submetidos ao modelo, alternando apenas os algoritmos. Na Figura 7.13 são apresentados os gráficos com as métricas precisão, revocação e F1.

A Figura 7.14 apresenta as matrizes de confusão dos algoritmos. As matrizes mostram que ao utilizar somente emojis, os algoritmos obtem melhor resultado, a maioria dos algoritmos no entanto, não mantem uma eficácia de acerto por nota estável, por exemplo, o SVM tem uma precisão de acerto de nota 1 de $\mathbf{92\% \pm 0{,}015}$ e de nota -1 $\mathbf{82\% \pm 0{,}018}$.

Na Figura 7.15 é feita a comparação entre os coeficientes de Mathew resultado da execução de cada algoritmo. Somente com o uso de emojis, o ganho dos coeficientes de Mathew foram em média $0,08$ com relação ao cenário de bigramas com uso somente de palavras. Assim sendo, o uso de emojis se mostrou mais assertivo para os modelos apresentados.

7.2.3 Usando Palavras + Emojis

Neste experimentos, foram combinadas palavras e os emojis transcritos. Na Figura 7.16 são apresentados os gráficos com as métricas precisão, revocação e F1.

A Figura 7.17 apresenta as matrizes de confusão dos algoritmos. As matrizes mostram que ao utilizar emojis com palavras, os algoritmos obtiveram os melhores resultados, a maioria dos algoritmos conseguiu manter uma estabilidade em relação à acurácia das avaliações de ambas as notas. O SVM foi o algoritmo que obteve melhor resultado, com acurácia geral de **93,0% ± 1,5**

Na Figura 7.18 é feita a comparação entre os coeficientes de Mathew resultado da execução de cada algoritmo. O gráfico demonstra uma melhora de $0,09$ em relação ao uso de unigramas somente com emojis. Desta forma, a combinação de palavras e emojis aumentou a eficácia do algoritmo, ou seja, ele foi mais assertivo na previsão das avaliações de ambas as notas. O SVM foi o mais assertivo, tanto na métrica F1, quanto no coeficiente de Mathew.

7.2.4 Resumo dos resultados

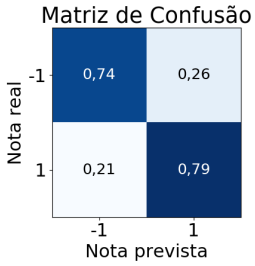
Na Tabela 7.2 são apresentados os resumos dos resultados dos gráficos, separados por algoritmo e estratégia de uso dos emojis. O algoritmo SVM obteve o maior desempenho geral, onde a estratégia mais eficaz foi a combinação de palavras + emojis, atingindo a acurácia de **91,0% ± 0,1** para a métrica F1 de ambas as classes. Dentre as estratégias de uso, pode-se perceber que o uso de emojis facilita ainda mais o trabalho do classificador, dando um ganho de cerca de **3,0% ± 0,1** quando comparadas as estratégias somente palavras contra somente emojis, e quando comparado ao uso de palavras combinados com emojis, este ganho sobe para cerca de **9,0% ± 0,1**.

7.3 Comentários finais do capítulo

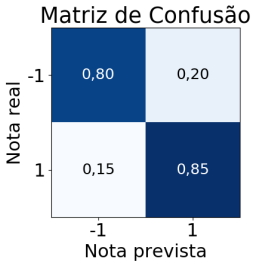
Neste capítulo foram apresentados os resultados dos experimentos. Os experimentos foram divididos entre o uso de unigramas e bigramas, e dentro de cada um, foram executados os modelos com três diferentes estratégias de utilização dos emojis, que foram: Somente palavras, somente emojis, e palavras combinadas com emojis. Com relação ao uso de unigramas e bigramas, percebeu-se que o uso de bigramas não melhorou a acurácia dos algoritmos, tendo resultados similares às execuções com uso de unigramas. Porém, o uso de bigramas aumentou o uso computacional, pois a bag-of-words continha duas palavras em cada posição, forçando o algoritmo a processar mais dados, o que aumenta a complexidade dos modelos. Em relação aos algoritmos, o que obteve melhor desempenho geral foi o SVM, assim como o estado da arte, em alguns cenários todavia, algoritmos como o Naiva Bayes e Random Forest conseguiram serem melhor

na identificação de avaliações de algumas notas, porém quando comparada a acurácia geral do modelo, o SVM se mostrou superior.

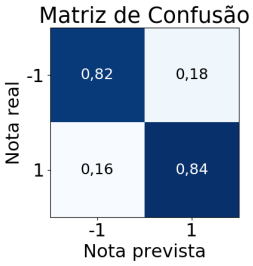
Com relação a estratégia de utilização dos emojis utilizando o SVM (que obteve melhores resultados), o uso de apenas palavras obteve uma acurácia geral média de **81,0% ± 0,020**, o uso de somente de emojis obteve **87,0% ± 0,013**, e a combinação de palavras e emojis obteve a acurácia geral média de **92,0% ± 0,012**. Desta forma, pode-se notar que o uso de emojis transcritos combinados com palavras (metodologia proposta) aumetou a acurácia do algoritmo em **9,0%**, ou seja, os emojis podem ser transcritos e têm uma contribuição significativa na classificação de polaridade.



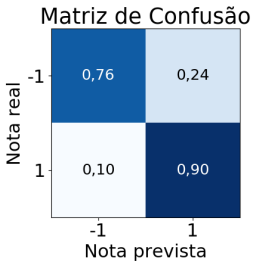
(a) Matriz Max Entropy.



(b) Matriz Naive Bayes.



(c) Matriz SVM.



(d) Matriz Random Forest.

Figura 7.2: Matrizes de confusão.

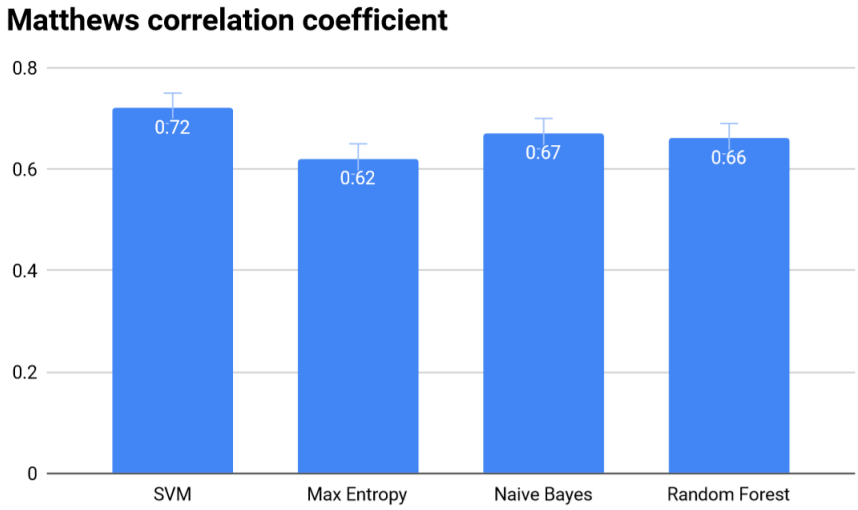
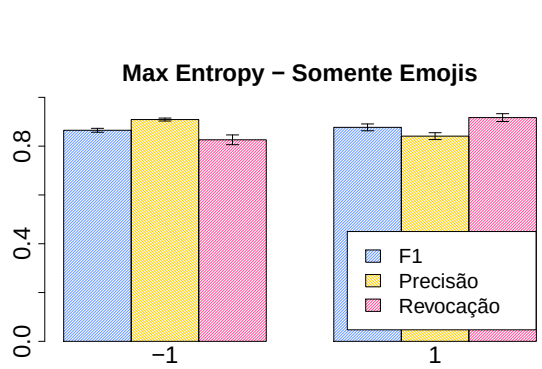


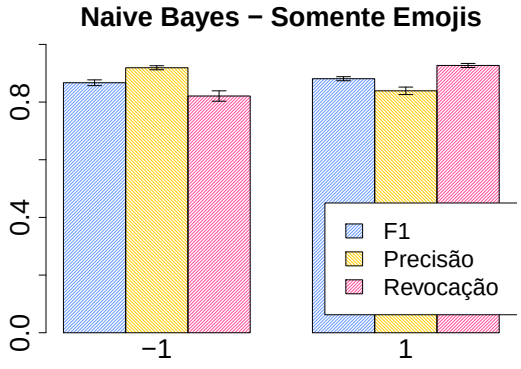
Figura 7.3: Comparação dos coeficientes de Mathew dos algoritmos.

Algoritmo	Nota	Estrategia de uso								
		Somente Palavras			Somente Emojis			Palavras + Emojis		
		Precisão	Revocação	F1	Precisão	Revocação	F1	Precisão	Revocação	F1
SVM	−1	**0,84 ± 0,019**	**0,81 ± 0,028**	**0,83 ± 0,018**	0,91 ± 0,013	0,81 ± 0,019	0,86 ± 0,009	**0,91 ± 0,016**	**0,90 ± 0,020**	**0,91 ± 0,013**
	1	**0,82 ± 0,023**	**0,84 ± 0,022**	**0,83 ± 0,016**	0,83 ± 0,013	0,92 ± 0,015	0,87 ± 0,006	**0,90 ± 0,018**	**0,91 ± 0,017**	**0,91 ± 0,012**
Random Forest	−1	0,89 ± 0,060	0,76 ± 0,060	0,82 ± 0,010	0,90 ± 0,040	0,82 ± 0,040	0,86 ± 0,000	0,91 ± 0,010	0,88 ± 0,002	0,90 ± 0,000
	1	0,79 ± 0,040	0,90 ± 0,070	0,84 ± 0,010	0,83 ± 0,040	0,91 ± 0,050	0,87 ± 0,010	0,89 ± 0,010	0,90 ± 0,000	0,90 ± 0,000
Naive Bayes	−1	0,84 ± 0,027	0,80 ± 0,028	0,82 ± 0,022	**0,91 ± 0,007**	**0,82 ± 0,018**	**0,86 ± 0,010**	0,88 ± 0,020	0,91 ± 0,019	0,89 ± 0,015
	1	0,81 ± 0,022	0,85 ± 0,028	0,83 ± 0,020	**0,91 ± 0,013**	**0,92 ± 0,007**	**0,88 ± 0,007**	0,90 ± 0,018	0,87 ± 0,022	0,89 ± 0,016
Max Entropy	−1	0,77 ± 0,024	0,73 ± 0,037	0,75 ± 0,027	0,90 ± 0,014	0,82 ± 0,020	0,86 ± 0,008	0,83 ± 0,019	0,81 ± 0,022	0,82 ± 0,022
	1	0,75 ± 0,027	0,79 ± 0,024	0,76 ± 0,021	0,84 ± 0,014	0,91 ± 0,016	0,87 ± 0,006	0,82 ± 0,018	0,83 ± 0,021	0,83 ± 0,015

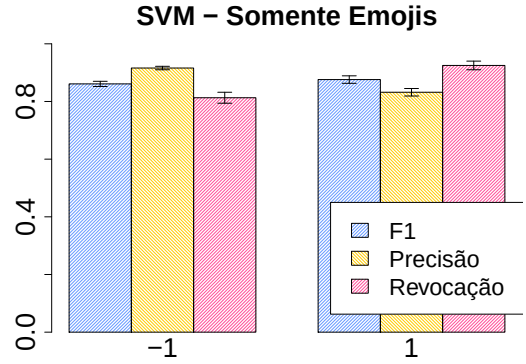
Tabela 7.1: Resumo de resultados com uso de unigramas.



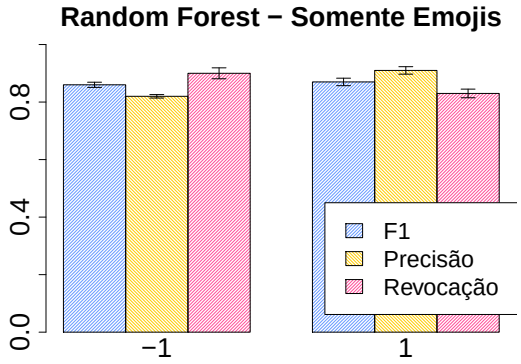
(a) Max Entropy.



(b) Naive Bayes.



(c) SVM.

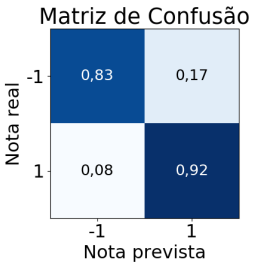


(d) Random Forest.

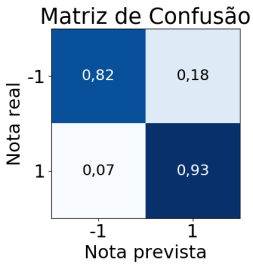
Figura 7.4: Resultados dos Algoritmos ao usar somente emojis.

Algoritmo	Nota	Estrategia de uso								
		Somente Palavras			Somente Emojis			Palavras + Emojis		
		Precisão	Revocação	F1	Precisão	Revocação	F1	Precisão	Revocação	F1
SVM	−1	**0,84 ± 0,019**	**0,82 ± 0,027**	**0,83 ± 0,018**	0,91 ± 0,013	0,82 ± 0,020	0,86 ± 0,009	**0,91 ± 0,015**	**0,90 ± 0,020**	**0,90 ± 0,013**
	1	**0,83 ± 0,022**	**0,84 ± 0,023**	**0,83 ± 0,016**	0,83 ± 0,014	0,91 ± 0,019	0,87 ± 0,007	**0,90 ± 0,018**	**0,91 ± 0,016**	**0,90 ± 0,012**
Random Forest	−1	0,82 ± 0,018	0,82 ± 0,026	0,82 ± 0,018	0,90 ± 0,011	0,82 ± 0,018	0,86 ± 0,009	0,89 ± 0,021	0,91 ± 0,019	0,89 ± 0,015
	1	0,82 ± 0,017	0,79 ± 0,031	0,83 ± 0,022	0,84 ± 0,013	0,90 ± 0,012	0,87 ± 0,007	0,89 ± 0,019	0,87 ± 0,024	0,89 ± 0,016
Naive Bayes	−1	0,86 ± 0,022	0,79 ± 0,031	0,83 ± 0,022	**0,92 ± 0,007**	**0,82 ± 0,018**	**0,86 ± 0,010**	0,89 ± 0,018	0,91 ± 0,019	0,90 ± 0,014
	1	0,81 ± 0,023	0,87 ± 0,022	0,84 ± 0,018	**0,83 ± 0,013**	**0,92 ± 0,007**	**0,88 ± 0,007**	0,90 ± 0,018	0,88 ± 0,020	0,89 ± 0,014
Max Entropy	−1	0,82 ± 0,021	0,78 ± 0,038	0,80 ± 0,023	0,88 ± 0,015	0,82 ± 0,018	0,85 ± 0,008	0,86 ± 0,014	0,86 ± 0,024	0,86 ± 0,014
	1	0,79 ± 0,028	0,82 ± 0,026	0,81 ± 0,018	0,83 ± 0,013	0,89 ± 0,017	0,86 ± 0,007	0,86 ± 0,021	0,86 ± 0,017	0,86 ± 0,012

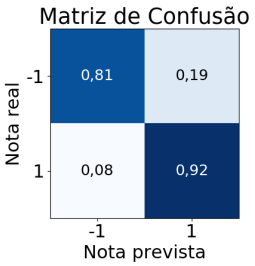
Tabela 7.2: Resumo de resultados com uso de bigramas.



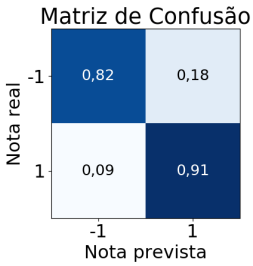
(a) Matriz Max Entropy.



(b) Matriz Naive Bayes.



(c) Matriz SVM.



(d) Matriz Random Forest.

Figura 7.5: Matrizes de confusão.

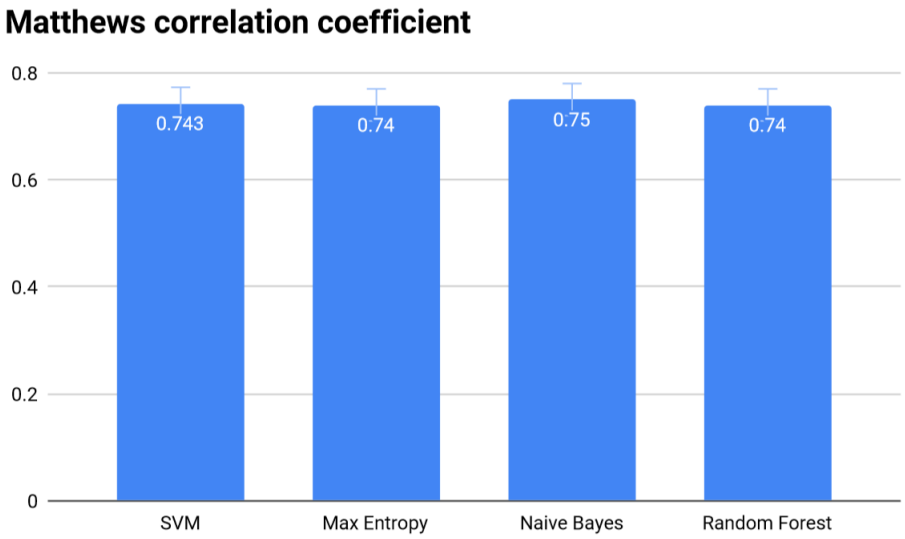
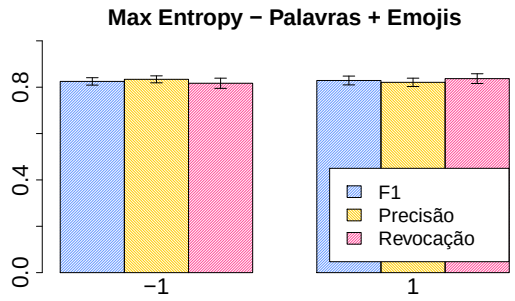
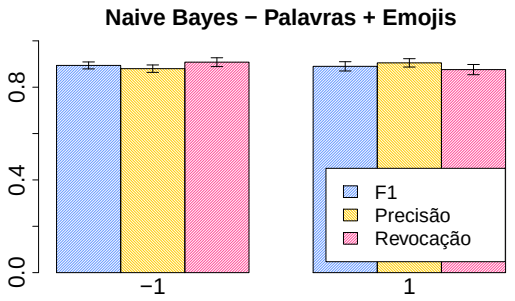


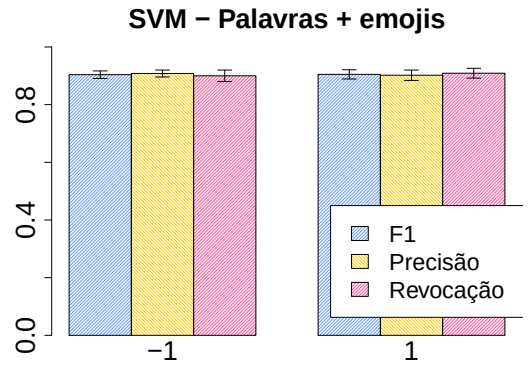
Figura 7.6: Comparação dos coeficientes de Mathew dos algoritmos.



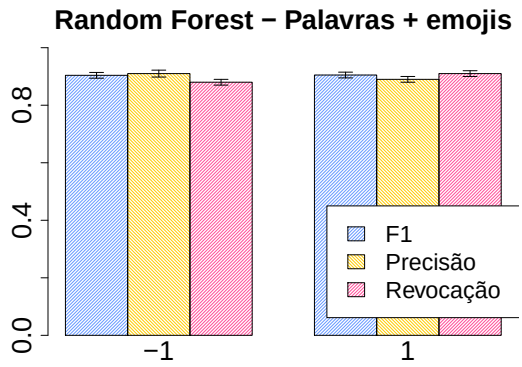
(a) Max Entropy.



(b) Naive Bayes.

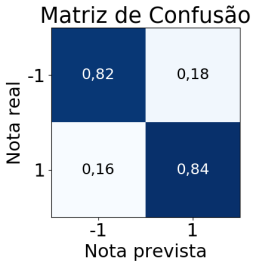


(c) SVM.

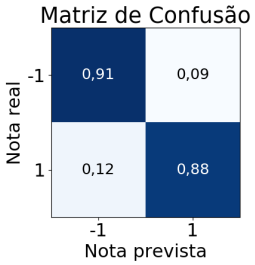


(d) Random Forest.

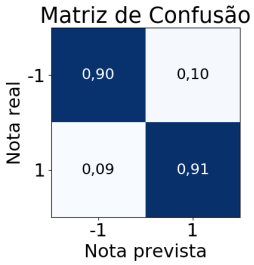
Figura 7.7: Resultados dos Algoritmos ao combinar palavras e emojis.



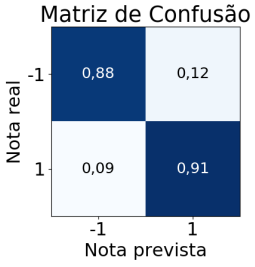
(a) Matriz Max Entropy.



(b) Matriz Naive Bayes.



(c) Matriz SVM.



(d) Matriz Random Forest.

Figura 7.8: Matrizes de confusão.

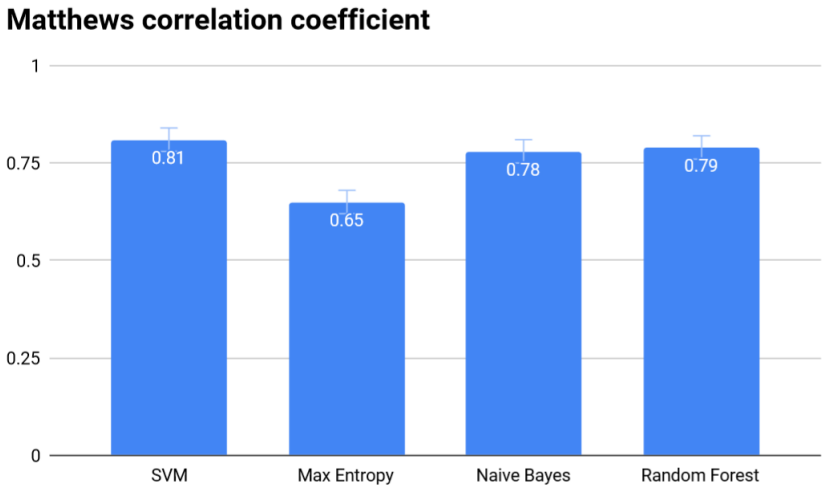
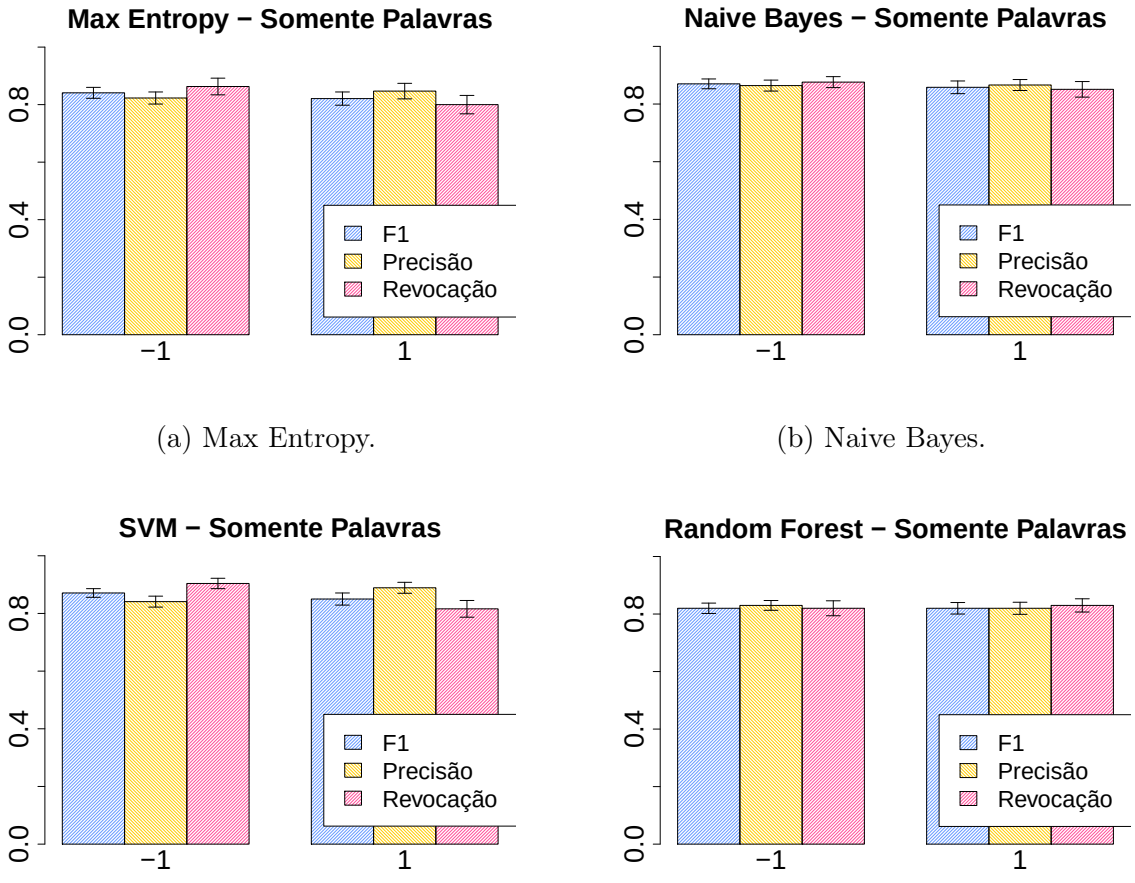


Figura 7.9: Comparação dos coeficientes de Mathew dos algoritmos.



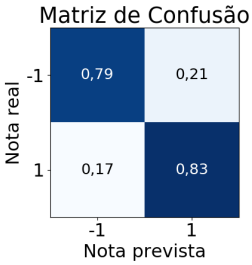
(a) Max Entropy.

(b) Naive Bayes.

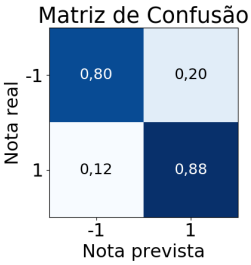
(c) SVM.

(d) Random Forest.

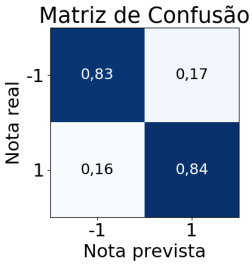
Figura 7.10: Resultados dos Algoritmos usando somente palavras.



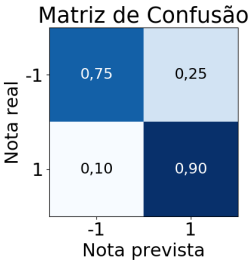
(a) Matriz Max Entropy.



(b) Matriz Naive Bayes.



(c) Matriz SVM.



(d) Matriz Random Forest.

Figura 7.11: Matrizes de confusão.

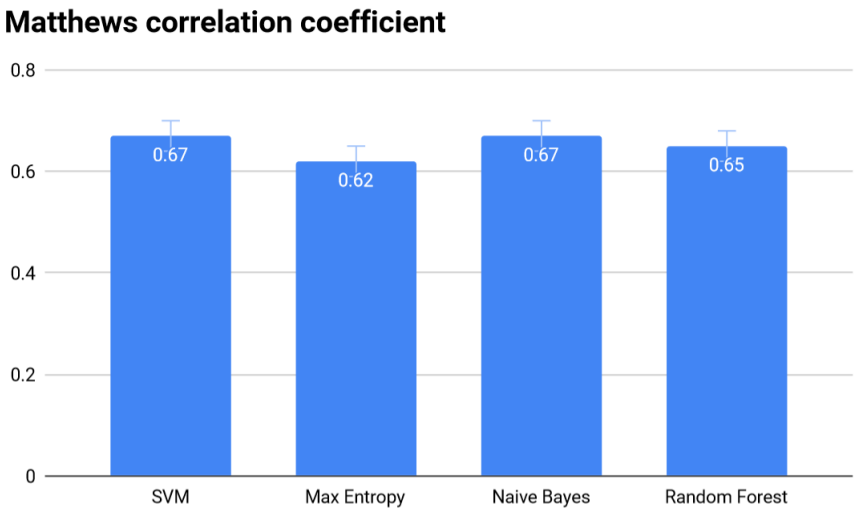
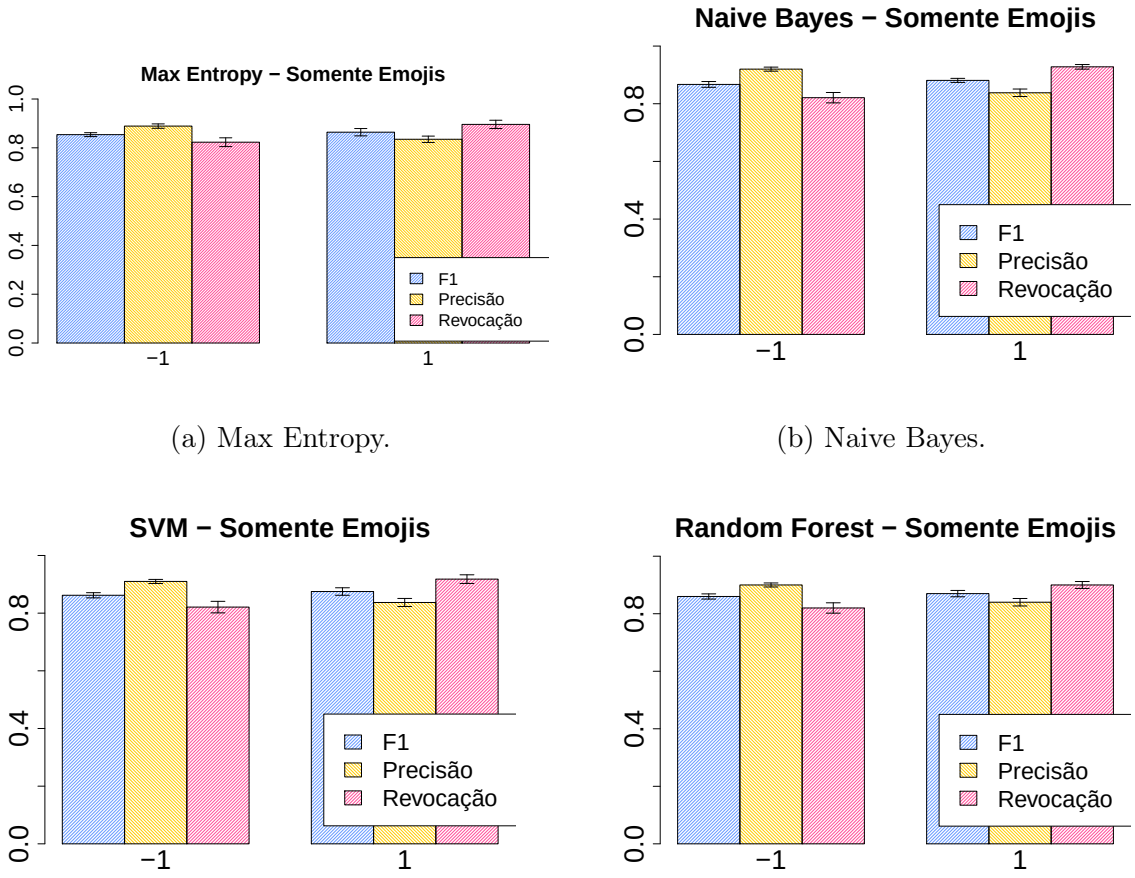


Figura 7.12: Comparação dos coeficientes de Mathew dos algoritmos.



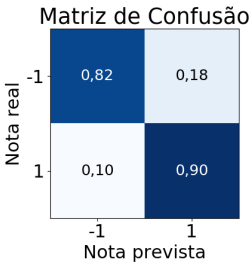
(a) Max Entropy.

(b) Naive Bayes.

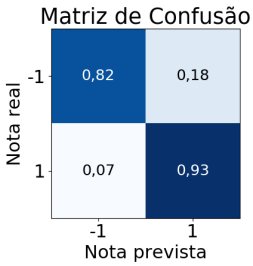
(c) SVM.

(d) Random Forest.

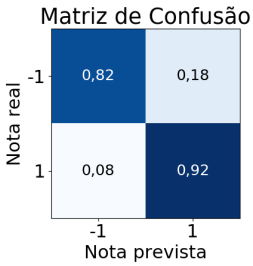
Figura 7.13: Resultados dos Algoritmos ao usar somente emojis.



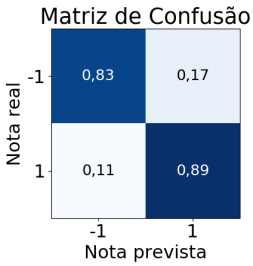
(a) Matriz Max Entropy.



(b) Matriz Naive Bayes.



(c) Matriz SVM.



(d) Matriz Random Forest.

Figura 7.14: Matrizes de confusão.

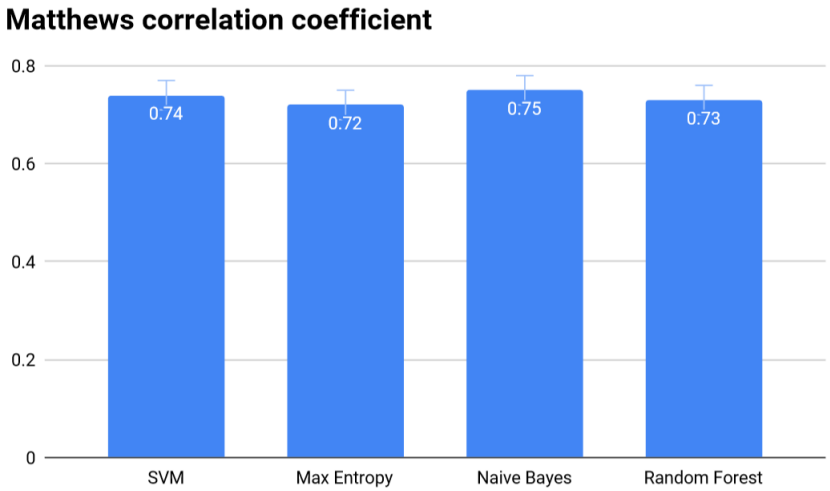
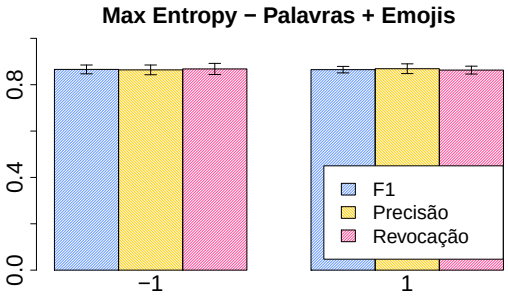
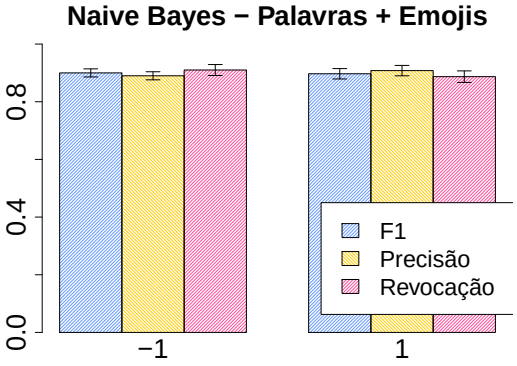


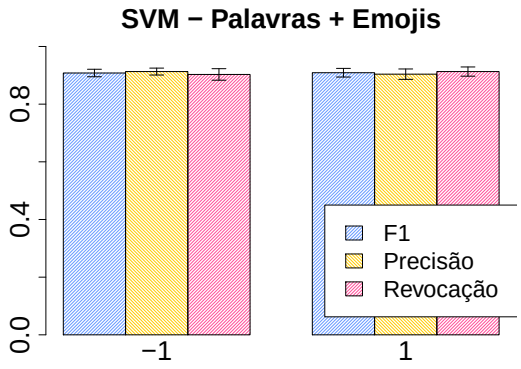
Figura 7.15: Comparação dos coeficientes de Mathew dos algoritmos.



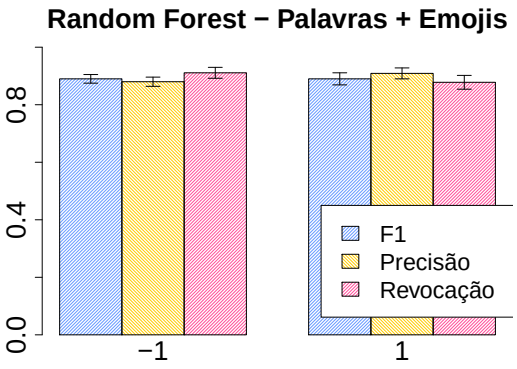
(a) Max Entropy.



(b) Naive Bayes.

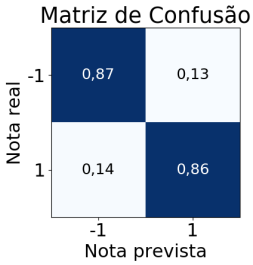


(c) SVM.

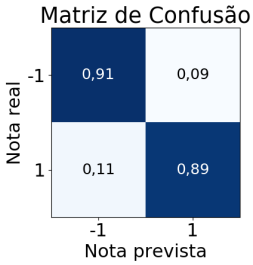


(d) Random Forest.

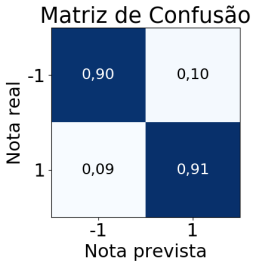
Figura 7.16: Resultados dos Algoritmos ao combinar palavras e emojis.



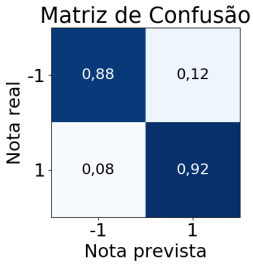
(a) Matriz Max Entropy.



(b) Matriz Naive Bayes.



(c) Matriz SVM.



(d) Matriz Random Forest.

Figura 7.17: Matrizes de confusão.

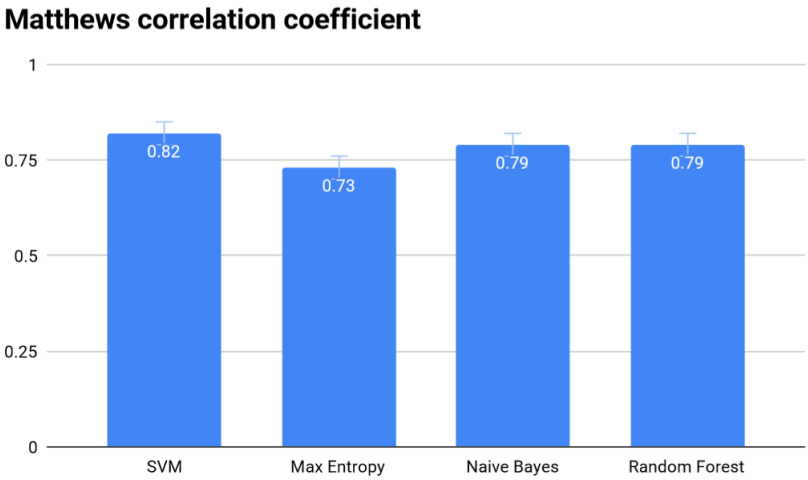


Figura 7.18: Comparação dos coeficientes de Mathew dos algoritmos.

Capítulo 8

Comentários Finais

8.1 Conclusão

Com a representação de características em unigramas e bigramas, podemos observar que o uso de unigramas se mostrou mais eficiente com relação ao tempo de execução e o desempenho obtido pelos algoritmos. Importante ressaltar que os resultados obtidos pelo uso de unigramas/bigramas para a representação utilizando somente emojis foi basicamente o mesmo, uma vez que a ordem de representação dos emojis não interfere no valor das notas previstas, o que não ocorre com as representações que utilizam palavras.

Dentre os tipos de utilização de emojis nas avaliações, as representações obtiveram resultados semelhantes, com uma leve superioridade da representação usando somente emojis, sendo que o experimento que tendeu a ser mais rápido é também utilizando somente os emojis, pois, a bag-of-words era menor, e havia menos a ser processado. A partir disso, para problemas com poucas classes (como o da polaridade), podemos inferir que o uso de emojis com BoW pode vir a ser mais eficiente (em termos de simplicidade e acurácia) que os métodos tradicionais.

Em relação aos classificadores utilizados, no geral, o que obteve o melhor desempenho em termos de acurácia foi o SVM (92 %). Importante ressaltar que somente o uso de emojis, o Naive Bayes obteve o melhor resultado (87%), o SVM obteve (86 %), porém o Naive Bayes é indicado para problemas com muitas classes, e por ser um algoritmo puramente estatístico baseado na probabilidade de Bayes, ele gasta muito recurso computacional, e portanto, o SVM seria o mais indicado para este cenário. Desta forma, podemos concluir que para este problema, assim como na literatura, o SVM é também o algoritmo com melhores resultados para a classificação de polaridade utilizando emojis.

Importante ressaltar que houve um pré-processamento nos emojis antes da execução dos algoritmos, foram transcrevidos todos os emojis em palavras únicas pré-definidas no apêndice A.

Portanto, emojis e emoticons são recursos linguísticos (linguagem online universal) com poder discriminante para previsão de opinião/polaridade/sentimento, principalmente, considerando opiniões extremas (muito negativas/positivas).

8.2 Limitações

O uso de emojis em ambientes virtuais online, apesar de ser muito utilizado, ainda não é extremamente utilizado em ambientes de lojas virtuais como a Google PlayStore, Amazon e outras. A quantidade de textos com emojis está diretamente relacionado com o dispositivo com o qual o usuário utiliza para a contrução do texto. Algumas avaliações foram feitas de computadores, onde não se tem o uso de emojis, e sim emoticons. Durante o pré-processamento dos textos, foram constatados alguns falso positivos, onde o usuário dava uma nota muito positiva, porém usava emojis/emoticons negativos e vice-versa. Isto está relacionado com o conceito de irônia, que é a utilização de palavras que manifestam o sentido oposto do seu significado literal. Desta forma, a ironia afirma o contrário daquilo que se quer dizer ou do que se pensa. Desta forma, a irônia foi utilizada a partir de emojis também. Além disso, algumas avaliações continham caracteres que pareciam ser emoticons pelo formato, mas eram partes de expressões e foram levados em consideração como emoticons.

Apesar dos resultados expressivos do modelo utilizando SVM, a maior fonte de erros eram notas que o modelo disse ser positiva, mas a nota era 2 (mais negativa), e da mesma forma para notas que ele disse ser negativa, porém era nota 4. Assim sendo, nota-se que avaliações mais interiores (2, 3 e 4) são mais difíceis de acertar pois o uso de emojis é bem diverso, ou seja, há emojis positivos e negativos juntos nestas avaliações. Uma coisa que podería ser aprimorada para tentar entender melhor este acontecimento, seria rodar um modelo que treine com notas 1 e 5, validando com notas 2 e 4, e classificando com notas 3. Desta forma, talvez pudessem ser obtidos resultados melhores com avaliações nota 3, onde geralmente fica concentrado avalições onde usuário usam de irônia e sarcasmo com abundância.

8.3 Trabalhos Futuros

Como a abordagem se mostrou boa na classificação em duas classes (positivo, e negativo), como trabalhos futuros, pretendemos modelar o problema com uma estrutura hierárquica de classificação, onde no primeiro nível de classificação as avaliações seriam classificadas em três classes (positivo, negativo e neutro). No segundo nível, iríamos classificar as avaliações negativas e positivas em outras duas classes cada, sendo elas: Muito negativa, pouco negativa, muito positiva e pouco positiva.

Desta forma, poderíamos usar a abordagem que obteve melhores resultados para especializar ainda mais as avaliações, usando dois níveis de classificação, primeiro nível teria três classes, e o segundo quatro classes (duas negativas e duas positivas). O número de amostras para a classificação hierárquica também será maior, com o objetivo de treinar melhor o modelo. A hierarquização esta melhor ilustrada na figura 8.1.

Além disso, outra área que poderia ser explorada seria a análise e classificação de sentimentos das avaliações com polaridades positivas e negativas, onde dado uma avaliação positiva, pudessem ser extráidos sentimentos que mais refletem o conteúdo da avaliação utilizando apenas emojis. Talvez a combinação de muitos emojis positivos e negativos numa única avaliação seja um caso de irônia ou sarcasmo. Caracterizar as combinações de emojis nas avaliações em relação a sentimentos pode trazer os mesmos benefícios de se classificar a polaridade, no sentido de que emojis sáo uma linguagem global e universal, ou seja, pode ser interpretado por todo o mundo.

Apêndice A

Tabela de Emojis

Emoji	Word
😀	emojixa
😁	emojixb
😂	emojixc
🤣	emojixd
😃	emojixd
😄	emojixe
😅	emojixf
😆	emojixg
😉	emojixh
😊	emojixi
😋	emojixj
😎	emojixk
😍	emojixl
😘	emojixm
😗	emojixn
😙	emojixo
😚	emojixp
☺️	emojixq
🙂	emojixr
🤗	emojixs
🤔	emojixt

Emoji	Word
🤨	emojixu
😐	emojixv
😑	emojixx
😶	emojixy
🙄	emojixz
😏	emojixaa
😣	emojixab
😥	emojixac
😮	emojixad
🤐	emojixae
😯	emojixaf
😪	emojixag
😫	emojixah
😴	emojixai
😌	emojixaj
😛	emojixak
😜	emojixal
😝	emojixam
🤤	emojixan
😒	emojixao
😓	emojixap

Emoji	Word
😔	emojixaq
😕	emojixar
🙃	emojixas
🤑	emojixat
😯	emojixau
☹	emojixav
😕	emojixax
😖	emojixay
😞	emojixaz
😟	emojixba
😤	emojixbb
😢	emojixbc
😭	emojixbd
😦	emojixbe
😧	emojixbf
😨	emojixbg
😩	emojixbh
🤯	emojixbi
😬	emojixbj
😰	emojixbk
😱	emojixbl

Emoji	Word
😳	emojixbm
😜	emojixbn
😵	emojixbo
😡	emojixbp
😠	emojixbq
🤬	emojixbr
😷	emojixbs
🤒	emojixbt
🤕	emojixbu
🤢	emojixbv
🤮	emojixbx
🤧	emojixbz
😇	emojixca
🤠	emojixcb
🤡	emojixcc
🥴	emojixcd
🤥	emojixce
🤭	emojixcf
🧐	emojixcg
🤓	emojixch
😈	emojixci

Emoji	Word
	emojixcj
	emojixck
	emojixcl
	emojixcm
	emojixcn
	emojixco
	emojixcp
	emojixcq
	emojixcr
	emojixcs
	emojixct
	emojixcu
	emojixcv
	emojixcx
	emojixcy
	emojixcz
	emojixda
	emojixdb
	emojixdc
	emojixdd
	emojixde

Emoji	Word
🙈	emojixdf
🙉	emojixdg
🙊	emojixdh
🐵	emojixdi
🐒	emojixdj
🦍	emojixdk
🦧	emojixdm
🐶	emojixdn
🐕	emojixdo
🦮	emojixdp
🐕‍🦺	emojixdq
🐩	emojixds
🐺	emojixdt
🦊	emojixdu
🦝	emojixdv
🐱	emojixdx
🐈	emojixdy
🐈‍⬛	emojixdz
🦁	emojixea
🐯	emojixeb
🐅	emojixec

Emoji	Word
👱🏼‍♀️	emojixed
👩🏽	emojixee
👩🏾	emojixef
👩🏿	emojixeg
👱‍♂️	emojixeh
👨🏻	emojixei
👱🏻‍♂️	emojixej
👨🏽	emojixek
👨🏾	emojixel
👨🏿	emojixem
👱‍♂️	emojixen
👨🏻	emojixeo
👱🏻‍♂️	emojixep
👨🏽	emojixeq
👨🏾	emojixeq
👨🏿	emojixer
👱‍♀️	emojixes
👩🏻	emojixet
👱🏻‍♀️	emojixeu
👩🏽	emojixev
👩🏾	emojixex

Emoji	Word
	emojixey
	emojixez
	emojixfa
	emojixfb
	emojixfc
	emojixfe
	emojixfg
	emojixfh
	emojixfi
	emojixfj
	emojixfk
	emojixfl
	emojixfm
	emojixfn
	emojixfo
	emojixfp
	emojixfq
	emojixfr
	emojixfs
	emojixft
	emojixfu

Emoji	Word
👈	emojixfv
👉	emojixfx
👊	emojixfy
✊	emojixfz
👎	emojixga
👍	emojixgb
👌	emojixgc
🖐	emojixgd
🖖	emojixge
🤘	emojixgf
🤚	emojixgg
🤞	emojixgh
✌	emojixgi
👇	emojixgj
☝	emojixgk
👆	emojixgl
🖕	emojixgm
👉	emojixgn
👈	emojixgo
💪	emojixgp
💋	emojixgq

Emoji	Word
💘	emojixgr
❤	emojixgs
💓	emojixgt
💔	emojixgu
💕	emojixgv
💖	emojixgx
💗	emojixgy
💙	emojixha
💚	emojixha
💛	emojixhb
🧡	emojixhc
💜	emojixhd
🖤	emojixhe
💝	emojixhf
💞	emojixhg
💟	emojixhh
❣	emojixhi
💌	emojixhj
💤	emojixhk
💦	emojixhl
💥	emojixhm

Emoji	Word
💢	emojixhn
💣	emojixho
🕶	emojixhp
🐵	emojixhq
🐒	emojixhr
🐶	emojixhs
🐋	emojixht
🐬	emojixhu
🌹	emojixhv
🌷	emojixhx
🙏	emojixzzz
♡	emojixaaa
:/	emoticonx
:\	emoticonx
'=\	emoticonx
'=/	emoticonx
/:	emoticonx
\:	emoticonx
\'=	emoticonx
/'=	emoticonx
:D	emoticonx

Emoji	Word
:d	emoticonx
:p	emoticonx
:P	emoticonx
'=D	emoticonx
'=d	emoticonx
'=p	emoticonx
'=P	emoticonx
'=s	emoticonx
'=S	emoticonx
:s	emoticonx
:S	emoticonx
'=x	emoticonx
'=X	emoticonx
:X	emoticonx
:x	emoticonx
:@	emoticonx
:b	emoticonx
:c	emoticonx
:C	emoticonx
:B	emoticonx
:3	emoticonx

Emoji	Word
'=3	emoticonx
'=]	emoticonx
:]	emoticonx
:-)	emoticonx
:-]	emoticonx
:-D	emoticonx
:-P	emoticonx
:-X	emoticonx
:-x	emoticonx
:-/	emoticonx
:-\	emoticonx
:-|	emoticonx
:-[	emoticonx
:-(	emoticonx
:-c	emoticonx
:-C	emoticonx
:-O	emoticonx
:-o	emoticonx
:-0	emoticonx
:0	emoticonx
:o	emoticonx

Emoji	Word
:O	emoticonx
;)	emoticonx
;D	emoticonx
;x	emoticonx
;X	emoticonx
;p	emoticonx
;P	emoticonx
;d	emoticonx
;*	emoticonx
:*	emoticonx
'=*	emoticonx
<3	emoticonx
s2	emoticonx
S2	emoticonx
SZ	emoticonx
sz	emoticonx
</3	emoticonx
</B	emoticonx
<|3	emoticonx
<|B	emoticonx
:)	emoticonx

Emoji	Word
:(	emoticonx
;(	emoticonx
:|	emoticonx
'=)	emoticonx
:'(	emoticonx
o_o	emoticonx
o_O	emoticonx
XD	emoticonx
:-@	emoticonx
:-V	emoticonx
:-3	emoticonx
:-$	emoticonx
:-!	emoticonx
:-Q	emoticonx
^^	emoticonx
^.^	emoticonx
^.o	emoticonx
o.^	emoticonx
o.o	emoticonx
O.O	emoticonx
O.o	emoticonx

Emoji	Word
o.O	emoticonx

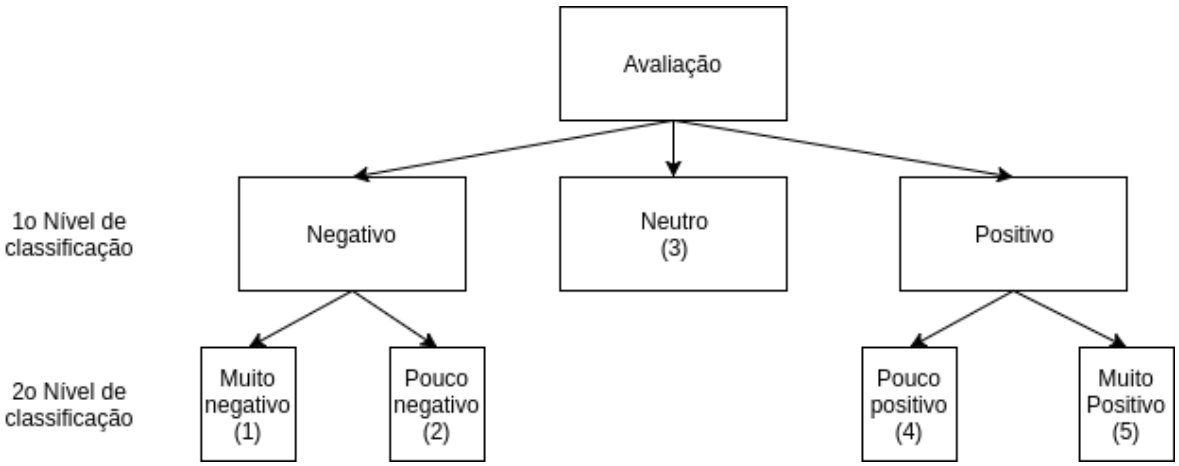


Figura 8.1: Classificação Hierarquica

Referências Bibliográficas

H. Abdi and L. J. Williams. Principal component analysis. *WIREs Comput. Stat.*, 2(4):433–459, July 2010. ISSN 1939-5108. doi: 10.1002/wics.101. URL `https://doi.org/10.1002/wics.101`.

F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd Conference on Hypertext and Social Media*, pages 187–196. ACM, 2012.

T. G. Almeida, B. A. Souza, A. A. Menezes, C. Figueiredo, and E. F. Nakamura. Sentiment analysis of portuguese comments from foursquare. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 355–358. ACM, 2016.

E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

S.-A. Bahrainian and A. Dengel. Sentiment analysis and summarization of twitter data. In *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*, pages 227–234. IEEE, 2013.

P. Barnaghi, P. Ghaffari, and J. G. Breslin. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 52–57. IEEE, 2016.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In D. D.

Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1471–1479. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6383-unifying-count-based-exploration-and-intrinsic-motivation.pdf.

K. Bhargava, T. Gujral, M. Chawla, and T. Gujral. Comment based seller trust model for e-commerce. In *Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on*, pages 387–391. IEEE, 2016.

S. K. Bharti, K. S. Babu, and S. K. Jena. Parsing-based sarcasm sentiment recognition in twitter data. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1373–1380. IEEE, 2015.

M. Bouazizi and T. O. Ohtsuki. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488, 2016.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

J. Burge. 5 Billion Emojis Sent Daily on Messenger. https://blog.emojipedia.org/5-billion-emojis-sent-daily-on-messenger/, 2018. [Online; accessed 04-June-2019].

E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

I. Csiszar. Maximum entropy and bayesian methods. *Kluwer Academic Publishers*, 2: 35–50, 1996. ISSN 1532-4435.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

D. Derks, A. E. Bos, and J. Von Grumbkow. Emoticons and social interaction on the internet: the importance of social context. *Computers in human behavior*, 23(1): 842–849, 2007.

S. Gangrade, N. Shrivastava, and J. Gangrade. Instagram sentiment analysis: Opinion mining. *SSRN Electronic Journal*, 01 2019. doi: 10.2139/ssrn.3372757.

P. Goncalves, F. Benevenuto, and V. Almeida. O que tweets contendo emoticons podem revelar sobre sentimentos coletivos. In *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2013.

C. Haythornthwaite. Social networks and internet connectivity effects. *Information, Communication & Society*, 8(2):125–147, 2005. doi: 10.1080/13691180500146185. URL https://doi.org/10.1080/13691180500146185.

W. A. Hussien, Y. M. Tashtoush, M. Al-Ayyoub, and M. N. Al-Kabi. Are emoticons good enough to train emotion classifiers of arabic tweets? In *Computer Science and Information Technology (CSIT), 2016 7th International Conference on*, pages 1–6. IEEE, 2016.

A. Illendula and A. Sheth. Multimodal emotion classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 439–449, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3316549. URL http://doi.acm.org/10.1145/3308560.3316549.

I. Inc. Instagram Live anouncement. https://canaltech.com.br/redes-sociais/instagram-bate-marca-de-1-bilhao-de-usuarios-ativos-116344/, 2018. [Online; accessed 19-May-2018].

N. Inc. Tchau estrelas, oi polegares! | Netflix. https://www.youtube.com/watch?v=MuDJeW16LwM, 2017. [Online; accessed 04-June-2019].

T. Inc. Some facts of Twitter. https://about.twitter.com/company, 2016. [Online; accessed 12-Dec-2016].

A. Ip. The impact of emoticons on affect interpretation in instant messaging, 2002, 2012.

N. Jindal and B. Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 244–251, New York, NY, USA, 2006a. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148215. URL http://doi.acm.org/10.1145/1148170.1148215.

N. Jindal and B. Liu. Mining comparative sentences and relations. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1331–1336. AAAI Press, 2006b. ISBN 978-1-57735-281-5. URL http://dl.acm.org/citation.cfm?id=1597348.1597400.

H. Kang, S. J. Yoo, and D. Han. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39 (5):6000–6010, 2012.

B. Kégl. The return of adaboost.mh: multi-class hamming trees. *CoRR*, abs/1312.6086, 2014.

E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades. Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, 40(10):4065–4074, 2013.

E. Kouloumpis, T. Wilson, and J. D. Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

S. Li. Sentiment classification using subjective and objective views. *International Journal of Computer Applications*, 80(7), 2013.

A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2 (3):18–22, 2002a. URL `http://CRAN.R-project.org/doc/Rnews/`.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002b. URL `https://CRAN.R-project.org/doc/Rnews/`.

B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

I. Marketing. Utilize uma lista de emojis na sua campanha de marketing e aumente a empatia dos clientes com a sua marca. `https://www.idealmarketing.com.br/blog/lista-de-emojis/`, 2019. [Online; accessed 04-June-2019].

B. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442 – 451, 1975. ISSN 0005-2795. doi: https://doi.org/10.1016/0005-2795(75)90109-9. URL `http://www.sciencedirect.com/science/article/pii/0005279575901099`.

W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–387. ACM, 2012.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-908-1. doi: 10.1145/1298306.1298311. URL `http://doi.acm.org/10.1145/1298306.1298311`.

A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1):93–107, 2014.

F. Moraes, M. Vasconcelos, P. Prado, J. Almeida, and M. Gonçalves. Polarity analysis of micro reviews in foursquare. In *Proceedings of the 19th Brazilian symposium on Multimedia and the web (WebMedia)*, pages 113–120. ACM, 2013.

M. Nakano. Why and how i created emoji: Interview with shigetaka kurita, 2016.

M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2): 167–256, 2003.

K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67, 1999.

B. A. Ojokoh and O. Kayode. A feature-opinion extraction approach to opinion mining. *Journal of Web engineering*, 11(1):51–63, 2012.

V. L. A. M. d. O. e. Paiva. A LINGUAGEM DOS EMOJIS. *Trabalhos em LinguÃstica Aplicada*, 55:379 – 401, 08 2016. ISSN 0103-1813. URL `http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-18132016000200379&nrm=iso`.

A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

L. Pardal and E. S. Lopes. *Métodos e técnicas de investiga-cão social*. Areal, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Prado. Netflix troca estrelas por polegares em novo sistema de avaliação. `https://tecnoblog.net/212159/netflix-avaliacao-polegares/`, 2018. [Online; accessed 04-June-2019].

E. Raisi and B. Huang. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 409–416, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4993-2. doi: 10.1145/3110025.3110049. URL `http://doi.acm.org/10.1145/3110025.3110049`.

J. D. Rennie, L. Shih, J. Teevan, D. R. Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, volume 3, pages 616–623. Washington DC), 2003.

A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12, 2012.

L. G. S. Selvan and T.-S. Moh. A framework for fast-feedback opinion mining on twitter data streams. In *Collaboration Technologies and Systems (CTS), 2015 International Conference on*, pages 314–318. IEEE, 2015.

J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma. Sentiment analysis. *Inf. Sci.*, 311(C):18–38, Aug. 2015a. ISSN 0020-0255. doi: 10.1016/j.ins.2015.03.040. URL `http://dx.doi.org/10.1016/j.ins.2015.03.040`.

J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma. Sentiment analysis: a review and comparative analysis of web services. *Information Sciences*, 311:18–38, 2015b.

S. Shah, K. Kumar, and R. K. Saravanaguru. Sentimental analysis of twitter data using classifier algorithms. *International Journal of Electrical and Computer Engineering*, 6(1):357–366, 2016a. ISSN 20888708. doi: 10.11591/ijece.v6i1.8982.

S. Shah, K. Kumar, and R. K. Sarvananguru. Sentimental analysis of twitter data using classifier algorithms. *International Journal of Electrical and Computer Engineering (IJECE)*, 6(1):357–366, 2016b.

R. Stein, P. Jaques, and J. Valiati. An analysis of hierarchical text classification using word embeddings. *Inf. Sci.*, 471:216–232, 01 2019. doi: 10.1016/j.ins.2018.09.001.

H. Sui, Y. Jianping, Z. Hongxian, and Z. Wei. Sentiment analysis of chinese micro-blog using semantic sentiment space model. In *Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on*, pages 1443–1447. IEEE, 2012.

D. Terrana, A. Augello, and G. Pilato. Automatic unsupervised polarity detection on a twitter data stream. In *Semantic Computing (ICSC), 2014 IEEE International Conference on*, pages 128–134. IEEE, 2014.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, Aug 1998. ISSN 0162-8828. doi: 10.1109/34.709601.

V. Vapnik. The nature of statistical learning theory, 1995.

G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratique. *Journal für die Reine und Angewandte Mathematik*, 133(3):97–178, 1908.

A. H. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE, 2010.

C. Wang, J. Lu, and G. Zhang. A semantic classification approach for online product reviews. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 276–279. IEEE Computer Society, 2005.

Z. Wang, X. Cui, L. Gao, Q. Yin, L. Ke, and S. Zhang. A hybrid model of sentimental entity recognition on mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016(1):253, 2016.

L. Wasden. Internet lingo dictionary: A parent's guide to codes used in chat rooms, instant messaging, text messaging, and blogs. *Office of the Attorney General, State of Idaho. Retrieved June*, 1:2008, 2006.

J. Waters, N. Weed, T. Bakken, N. Graddis, N. Gouwens, D. Millman, and M. Hawrylycz. Identification of genetic markers for cortical areas using a random forest classification routine and the allen mouse brain atlas, 02 2019.

B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite. Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual review of sociology*, pages 213–238, 1996.

T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

L. K. Wives. *Descobrindo eventos locais utilizando análise de séries temporais nos dados do Twitter*. PhD thesis, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2013.

Q. Ye, Z. Zhang, and R. Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3):6527–6535, 2009.

L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Know.-Based Syst.*, 41:89–97, Mar. 2013. ISSN 0950-7051. doi: 10.1016/j.knosys.2013.01.001. URL `http://dx.doi.org/10.1016/j.knosys.2013.01.001`.