UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MARCOS PAULO MAIA DOS SANTOS

CLASSIFICAÇÃO AUTOMÁTICA DE FACETAS DE USABILIDADE E EXPERIÊNCIA DE USUÁRIO EM POSTAGENS RELACIONADAS AO USO

RUSSAS

2018

MARCOS PAULO MAIA DOS SANTOS

CLASSIFICAÇÃO AUTOMÁTICA DE FACETAS DE USABILIDADE E EXPERIÊNCIA DE USUÁRIO EM POSTAGENS RELACIONADAS AO USO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Alexandre Matos Arruda
Co-Orientadora: Profa. Dra. Marília Soares Mendes

RUSSAS

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S236c Santos, Marcos Paulo Maia dos.
Classificação automática de facetas de Usabilidade e Experiência de Usuário em Postagens Relacionadas ao Uso / Marcos Paulo Maia dos Santos. – 2018.
51 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2018.
Orientação: Prof. Dr. Alexandre Matos Arruda.
Coorientação: Profa. Dra. Marília Soares Mendes.

1. Postagens Relacionadas ao Uso. 2. Usabilidade e Experiência do Usuário. 3. Classificação Textual. 4. Machine Learning. 5. Deep Learning. I. Título.

CDD 005

MARCOS PAULO MAIA DOS SANTOS

CLASSIFICAÇÃO AUTOMÁTICA DE FACETAS DE USABILIDADE E EXPERIÊNCIA DE USUÁRIO EM POSTAGENS RELACIONADAS AO USO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Alexandre Matos Arruda (Orientador)
Universidade Federal do Ceará (UFC)

Profa. Dra. Marília Soares Mendes (Co-Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Ms. Tatiane Fernandes Figueiredo
Universidade Federal do Ceará (UFC)

Prof. Ms. Alex Lima Silva
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Agradeço à Deus, por ter me dado a força suficiente nessa jornada, ao meus pais por fazerem todo o possível para garantir aos filhos a melhor educação, sempre dando o maior apoio e incentivo. Ao meu professor orientador Alexandre Matos Arruda e professora coorientadora Marília Soares Mendes por todas as lições e apoio no desenvolvimento deste trabalho. À professora Maria Viviane de Menezes, minha primeira orientadora, por me possibilitar ingressar no mundo da pesquisa e por todos os ensinamentos e apoio durante a graduação. Aos meus tios, pelo imenso acolhimento e conforto que me propiciaram durante toda a minha estadia em Russas. À minha namorada, Tainá Lima Meneses, por todo o carinho, atenção, compreensão nos momentos difíceis e por me incentivar tanto. Aos meus colegas Afonso e Paloma, com os quais enfrentei grandes desafios, e aos colegas de classe por todos os momentos de diversão e cooperação durante o curso. Aos maravilhosos professores que colaboraram para a minha formação, à Universidade Federal do Ceará (UFC), ao Laboratório INterdisciplinar de Computação e Engenharia de Software (LINCE), ao Núcleo de Estudos em Machine Learning e Otimização (NEMO) e a todos que contribuíram de forma direta ou indireta a este trabalho e a minha graduação.

"There's no such thing as a painless lesson, they just don't exist. Sacrifices are necessary. You can't gain anything without losing something first. Although if you can endure that pain and walk away from it, you'll find that you now have a heart strong enough to overcome any obstacle. Yeah... a heart made fullmetal."

(Edward Elric)

RESUMO

Os feedbacks dos usuários são muito importantes em Interação Humano-Computador, para termos conhecimento sobre a Usabilidade e Experiência do Usuário (UUX) de um sistema. Recentemente, narrativas na forma de texto, expressas espontaneamente em sistemas sociais, lojas de aplicativos ou sites de revisões de produtos, denominadas Postagens Relacionadas ao Uso (PRU), têm se mostrado uma fonte valiosa de informações sobre a qualidade de uso. Identificar as metas de UUX (eg. satisfação, eficiência) em texto é, em essência, um problema de classificação em categorias, e tem sido realizada principalmente de forma manual. Tal tipo de tarefa é um dos muitos cujo estado-da-arte foi superado atualmente por uma classe de técnicas de aprendizado de máquina chamada Deep Learning. Este trabalho visa aplicar tais técnicas do estado-da-arte, além de técnicas mais tradicionais de aprendizado, para classificar metas de UUX em PRUs.

Palavras-chave: Postagens Relacionadas ao Uso. Usabilidade e Experiência do Usuário. Machine Learning. Deep Learning. Classificação Textual.

ABSTRACT

Users‘ feedbacks are very important in Human-Computer Interaction, for we to have knowledge regarding to a system’s Usability and User Experience (UUX). Recently, narratives in the form of text, expressed spontaneously in social systems, app stores or review sites, named Postings Related to Use (PRU), has been showed themselves as precious sources of information about the quality of use. Identifying UUX facets (eg. satisfaction, efficiency) in text is, in essence, a problem of classification into categories, and that has been mainly held in a manual way. Such kind of task is one of the many whose state-of-the-art has been currently overcamed by a class of machine learning techniques called Deep Learning. This research aims to apply these state-of-the-art algorithms, besides more traditional learning techniques, in order to classify UUX goals on PRUs.

Keywords: Posting Related to Use. Usability and User Experience. Machine Learning. Deep Learning. Text Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CNN	Convolutional Neural Network
IHC	Interação Humano-Computador
MLP	Multilayer Perceptron
NB	Naive Bayes
NLTK	Natural Language Toolkit
PRU	Postagens Relacionadas ao Uso
RNN	Recurrent Neural Network
RSLP	Removedor de Sufixos da Lingua Portuguesa
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TF-IDF	Term Frequency & Inverse Document Frequency
UUX	Usabilidade e Experiência do Usuário
UX	Experiência do Usuário

SUMÁRIO

1 INTRODUÇÃO

Recentemente, na área de Interação Humano-Computador (IHC), vários trabalhos têm buscado estudar as narrativas dos usuários sobre um sistema, para entender ou mesmo avaliar a Usabilidade e Experiência do Usuário (UUX) em relação a ele (KORHONEN *et al.*, 2010; OLSSON; SALO, 2012; MENDES *et al.*, 2013; HEDEGAARD; SIMONSEN, 2013; MENDES *et al.*, 2014; MENDES, 2015; MENDES *et al.*, 2015; FREITAS *et al.*, 2016; LIMA *et al.*, 2017; SILVA *et al.*, 2017; MENDES; FURTADO, 2017). Essas narrativas, denominadas Postagens Relacionadas ao Uso (PRU), segundo a metodologia MALTU (MENDES, 2015), podem dizer muito sobre as dificuldades enfrentadas pelos usuários no uso e sobre como eles se sentem. Elas podem ser encontradas em vários ambientes como em seção de avaliação de *App Stores*, em *feed* de redes sociais, em fóruns de sistemas sociais.

Dos estudos acima, somente os de Hedegaard & Simonsen (2013) e Mendes & Furtado (2017) tinham como objetivo usar o computador para auxiliar na tarefa de identificar aspectos de UUX nas postagens, e acabaram não atingindo altos níveis de qualidade de classificação, devido à característica muito subjetiva da extração de informações sobre UUX de textos. Mendes & Furtado (2017) consideram postagens na língua portuguesa e Hedegaard & Simonsen (2013) lidam com a língua inglesa. Outras pesquisas usaram a abordagem da classificação manual, sendo indicada como a mais precisa encontrada pelos pesquisadores, porém exigindo grande esforço. É relatado o quanto é complicado classificar manualmente as facetas de UUX em PRUs, tendo que ser feita por especialistas experientes e por mais de uma pessoa, para validar a análise. Essa dificuldade implica na impossibilidade de classificar um grande número de postagens, comparado ao volume de dados gerado intermitentemente pelos usuários. Freitas et al. (2016), Silva et al. (2017), Mendes & Furtado (2017) ainda enfrentam o problema de lidar com a ambiguidade e nuances da complexa língua portuguesa.

O problema de classificação textual tem sido abordado há muito tempo, por meio de técnicas de processamento de linguagem natural, mineração de dados, e, mais recentemente, por classificadores construídos com algoritmos de aprendizado de máquina (*machine learning*). O rápido crescimento de *machine learning* tem revolucionado muitas áreas dentro e fora da ciência da computação, auxiliando na resolução de problemas de classificação, predição e clusterização. Os principais grandes avanços na área de aprendizado têm surgido de uma classe de técnicas chamada *Deep Learning* (aprendizado profundo), onde se tem arquiteturas baseadas fortemente em redes neurais (LECUN *et al.*, 2015). Esses métodos conseguem, a partir da entrada de

dados brutos, extrair características automaticamente, e obter um alto nível de abstração após várias transformações aplicadas aos dados de entrada, possibilitando o aprendizado de funções muito complexas. Para classificação de texto, cada vez mais *Deep Learning* vem mostrando surpreendentes resultados frente aos métodos mais convencionais de aprendizado de máquina (KIM, 2014; ZHANG *et al.*, 2015; LAI *et al.*, 2015; CONNEAU *et al.*, 2017).

Diante do contexto apresentado, a proposta deste trabalho é classificar de forma automática postagens em português relacionadas ao uso de acordo com as facetas de usabilidade e experiência do usuário (listadas no capítulo de Fundamentação Teórica) por meio de técnicas de classificação do estado-da-arte da inteligência artificial. Os resultados podem ser usados para aprimorar a ferramenta UUX-Posts (MENDES; FURTADO, 2017) e para auxiliar praticantes de IHC, sejam pesquisadores ou do mercado, que desejarem investigar a UUX em postagens de usuário.

1.1 Objetivos

1.1.1 Objetivo geral

Desenvolver um modelo para identificação automática de facetas de Usabilidade e Experiência de Usuário em Postagens Relacionadas ao Uso por meio de técnicas do estado-da-arte de classificação textual.

1.1.2 Objetivo específicos

- Investigar quais métodos automáticos de classificação lidam melhor com rótulos em geral muito subjetivos;
- Analisar correlação entre as facetas, padrões de identificação e heurísticas para melhor classificação;
- Comparar precisão, abrangência e qualidade de abordagens anteriores.

1.2 Organização do Trabalho

O trabalho está dividido em 6 capítulos. No Capítulo 2, são apresentados conceitos sobre IHC e *Machine Learning* necessários para o entendimento da pesquisa realizada. O Capítulo 3 contém alguns trabalhos relacionados ao tema da monografia. No Capítulo 4, é

descrita a metodologia seguida para estudar o problema e classificar as PRUs de acordo com as facetas de UUX. O Capítulo 5 apresenta os resultados obtidos no trabalho e o Capítulo 6 contém as considerações finais e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Postagens Relacionadas ao Uso (PRU)

PRU foi um termo criado para denominar artefatos textuais produzidos por usuários de um sistema que dizem algo sobre ele. Essas narrativas podem ser, por exemplo, reviews de produtos em lojas ou sites de reclamações, posts em redes sociais (lugar que é intrinsecamente recheado de PRUs (MENDES *et al.*, 2014)) ou em fóruns de outros sistemas sociais, como sistemas acadêmicos. Um exemplo de PRU extraída do Twitter: "gente esse twitter novo eh horriver não gostei".

Segundo a metodologia MALTU (MENDES, 2015), essas postagens podem ser classificadas em diversas categorias. Por tipo: se é dúvida, elogio, crítica, etc.; por polaridade: se é positiva, negativa ou neutra; por intenção: se é visceral, comportamental ou reflexiva; e por critério de qualidade de uso, no caso metas de Usabilidade e Experiência do Usuário: se fala de satisfação, eficácia, utilidade, afeto, frustração, etc.

2.2 Facetas de Usabilidade e Experiência do Usuário

Segundo a ISO 9241-210 (2010), Usabilidade e Experiência do Usuário dizem respeito à dimensão pela qual um produto, sistema ou serviço pode ser utilizado por usuários específicos para atingir objetivos específicos com eficácia, eficiência e satisfação em um contexto de uso específico, e à percepção de uma pessoa e às respostas resultantes do uso e/ou do uso antecipado de um produto, sistema ou serviço. Esses critérios de qualidade de uso são caracterizados por facetas (ou metas) que são dimensões, pontos de vista, fenômenos, ou aspectos indicadores de uma boa ou má qualidade, como por exemplo eficiência em Usabilidade (ISO, 2010) e motivação em Experiência do Usuário (UX) (BARGAS-AVILA; HORNBÆK, 2011).

Neste trabalho, foram consideradas as metas de UUX para classificar PRUs descritas no Quadro 1. Para o exemplo de PRU dado acima, é possível perceber que o usuário demonstra Frustração e se refere muito provavelmente à Estética do sistema.

Quadro 1 – Facetas de UUX

Metas	Característica	Referências
Eficácia	Alcançar os objetivos desejados durante o uso	(ISO, 2010; ROGERS *et al.*, 2013)
Eficiência	Recursos gastos em relação a execução de tarefas	(ISO, 2010; ROGERS *et al.*, 2013)
Segurança	Prevenir que o usuário cometa erros, e também se recupere de erros	(ROGERS *et al.*, 2013)
Utilidade	Fornecer funcionalidades úteis para que o usuário realize suas tarefas	(ROGERS *et al.*, 2013)
Memorabilidade	Ser fácil de lembrar como se usa	(ROGERS *et al.*, 2013)
Aprendizado	Ser fácil de aprender a usar	(ROGERS *et al.*, 2013)
Satisfação	Atitudes positivas e ausência de desconforto em relação ao uso do sistema	(ISO, 2010; ROGERS *et al.*, 2013)
Afeto	Afeto e emoção induzido pelo uso de um sistema	(BARGAS-AVILA; HORNBÆK, 2011)
Confiança	Medida em que os usuários estão satisfeitos que o sistema vai se comportar conforme o esperado	(BEVAN, 2008)
Estética	Apreciação da beleza ou bom gosto	(BARGAS-AVILA; HORNBÆK, 2011)
Frustação	Frustração ou sofrimento induzido pelo uso do sistema	(BARGAS-AVILA; HORNBÆK, 2011)
Motivação	Manter o usuário com vontade de usar o sistema	(BARGAS-AVILA; HORNBÆK, 2011)
Suporte	Apoio ao ser humano ou software de apoio disponíveis e como isso afeta a satisfação do usuário	(KETOLA; ROTO, 2008)

Fonte: O autor (2018).

2.3 Machine Learning

Machine Learning (Aprendizado de Máquina) é uma área da inteligência artificial que abrange os algoritmos capazes de reconhecer padrões e fazer predições ou agrupamentos. Dentre os principais tipos de aprendizagem, será considerado neste trabalho o *aprendizado supervisionado*, onde a máquina aprende uma função que mapeia entradas e saídas a partir da observação de alguns exemplos de pares de entradas e saídas (RUSSELL; NORVIG, 2016).

Neste trabalho, foram usados populares algoritmos de aprendizado de máquina, como Support Vector Machine (SVM), Naive Bayes (NB), classificador Stochastic Gradient Descent (SGD) e Random Forest.

2.3.1 Deep Learning

Deep Learning (Aprendizado Profundo) é uma subárea de *Machine Learning* que abrange modelos computacionais formados por muitas camadas de processamento que permitem obter um alto nível de abstração para representação dos dados.

Uma rede neural artificial simples consiste de muitas unidades de processamento conectadas, simulando os neurônios. Na arquitetura Multilayer Perceptron (MLP), há um conjunto de neurônios para receber uma entrada (sequência numérica), que passam para camadas

de neurônios conectadas a eles (denominadas camadas ocultas) até chegar a uma camada que fornecerá a saída (sequência numérica). Cada neurônio computa uma função que soma suas entradas multiplicadas por pesos, que vão definir a importância de cada entrada, e aplica uma função de ativação, que decide a relevância daquele neurônio. O treinamento de uma rede neural tem como objetivo ajustar os pesos associados às entradas de forma a minimizar uma função de erro que mede a distância entre as saídas da rede e as saídas desejadas (indicadas pela base de treinamento), por meio de um processo conhecido como *backpropagation*. Após treinar, a performance do algoritmo é medida em um conjunto diferente de exemplos, para verificar a habilidade de generalização da máquina. Espera-se que o sistema de aprendizado não se vicie em dar bons resultados somente com exemplos muito parecidos com os de treino (*overfitting*). (GOODFELLOW *et al.*, 2016)

Os dados de entrada para o aprendizado, são representados por *features*, que definem as características de um determinado dado. Por exemplo, para dados de frutas, as *features* podem ser cor, tamanho, peso, etc. Muitas das atuais aplicações de *machine learning* usam dados representados por *features* criadas por algoritmos externos ao de aprendizado (LECUN *et al.*, 2015). Essa representação é um problema importante em classificação de texto, geralmente baseada no modelo *bag-of-words* ou *unigrams*, *bi-grams*, *n-grams* ou algum outro método, os quais frequentemente ignoram informações como contexto ou ordem das palavras, não permitindo um aprendizado da semântica das palavras (LAI *et al.*, 2015). Com redes neurais profundas, é possível aprender representações a partir de texto bruto, assim como *features* de imagem, áudio e vários tipos de dados complicados de se extrair características.

2.3.1.1 Word embeddings

Word embeddings (vetores de palavras) são uma forma eficiente de prover representações significativas para palavras, onde palavras similares tenham representações similares. Eles são vetores de números reais que representam palavras em um espaço de *n* dimensões, capazes de capturar informações sintáticas, semânticas e morfológicas (HARTMANN *et al.*, 2017).

Um vetor é associado a uma palavra de forma que a distância entre dois vetores capture parte da relação semântica entre as palavras associadas, e podem ser aprendidos em uma camada de embeddings em uma rede neural. Eles podem ser gerados também por um algoritmo de aprendizado externo, sendo chamados de vetores pré-treinados.

2.3.1.2 Modelos

Neste trabalho, foram considerados vários modelos de *Deep Learning* atuais com recente destaque para classificação de texto, como redes neurais convolutivas — Convolutional Neural Network (CNN) (KIM, 2014) — , redes neurais recorrentes (LAI *et al.*, 2015) — Recurrent Neural Network (RNN) — com unidades LSTM (Long Short-Term Memory) e com unidades GRU (Gated Recurrent Unit), e as variacões LSTM com CNN e GRU com CNN.

2.4 Problema de Clasificação Multi-label

Em um problema de classificação *multi-label*, cada exemplo é associado com um conjunto de rótulos. No caso, uma Postagem Relacionada ao Uso pode estar relacionada à uma ou mais facetas de Usabilidade e Experiência do Usuário. Por exemplo, uma PRU "Muito bom, facilita muito na dieta"pode ser classificada como "Satisfação"e "Motivação".

Na literatura, há duas principais abordagens para múltiplos rótulos: a) transformação do problema de classificação *multi-label* em problemas de classificação *single-label*, para os quais podem ser usados muitos algoritmos conhecidos, e b) adaptação de algoritmos, para que possam lidar diretamente com múltiplos rótulos (TSOUMAKAS; KATAKIS, 2007).

2.4.1 Métodos de transformação do problema

2.4.1.1 Binary Relevance

O método *Binary Relevance* (Relevância Binária) basicamente trata cada rótulo como um problema de classificação isolado (ZHANG *et al.*, 2018). Isso significa que ele ignora qualquer relacionamento entre os rótulos. O Quadro 2 apresenta um exemplo da transformação de um problema usando a técnica *Binary Relevance*, sendo a base original exibida na esquerda e as transformações à sua direita.

Quadro 2 – Exemplo da aplicação de *Binary Relevance*.

X	y_1	y_2	y_3
x_1	1	0	1
x_2	0	1	1
x_3	0	0	1
x_4	1	1	0

X	y_1
x_1	1
x_2	0
x_3	0
x_4	1

X	y_2
x_1	0
x_2	1
x_3	0
x_4	1

X	y_3
x_1	1
x_2	1
x_3	1
x_4	0

Fonte: O autor (2018).

2.4.1.2 Classifier Chain

Na técnica *Classifier Chain* (Cadeia de Classificador) (READ *et al.*, 2011) também é treinado um classificador para cada rótulo, porém são adicionados os rótulos dos classificadores anteriores como entrada do algoritmo. De forma acumulativa, cada iteração considera uma quantidade maior de rótulos. Um exemplo da aplicação desse método é mostrado no Quadro 3, com a base original exibida na esquerda e as transformações à sua direita, separando com linha dupla os dados de entrada do rótulo associado.

Quadro 3 – Exemplo da aplicação de *Classifier Chain*.

X	y_1	y_2	y_3
x_1	1	0	1
x_2	0	1	1
x_3	0	0	1
x_4	1	1	0

X	y_1
x_1	1
x_2	0
x_3	0
x_4	1

X	y_1	y_2
x_1	1	0
x_2	0	1
x_3	0	0
x_4	1	1

X	y_1	y_2	y_3
x_1	1	0	1
x_2	0	1	1
x_3	0	0	1
x_4	1	1	0

Fonte: O autor (2018).

2.4.1.3 Label Powerset

Label Powerset (Conjunto das partes de Rótulos) é uma abordagem para transformação do problema que considera um novo rótulo para cada combinação única de rótulos, que é atribuído no lugar dos rótulos associados originalmente (TSOUMAKAS *et al.*, 2009). Como pode ser observado no exemplo do Quadro 4, o problema *multi-label* é transformado em apenas um problema *single-label*.

Quadro 4 – Exemplo da aplicação de *Label Powerset*.

X	y_1	y_2	y_3
x_1	1	0	1
x_2	0	1	1
x_3	0	0	1
x_4	1	1	0

X	y
x_1	{1, 0, 1}
x_2	{0, 1, 1}
x_3	{0, 0, 1}
x_4	{1, 1, 0}

Fonte: O autor (2018).

2.4.2 Métodos de adaptação de algoritmos

As abordagens por adaptação, como o nome obviamente sugere, modificam algoritmos conhecidos para realizar classificação *multi-label*, ao invés de criar outros conjuntos de problemas. Alguns desses algoritmos usam as ideias de transformação de problema em sua

construção.

O algoritmo ML-kNN (ZHANG; ZHOU, 2007) é uma adaptação de *k-Nearest Neighbours* que usa o algoritmo original independentemente para cada classe, encontrando os k exemplos mais próximos e selecionando rótulos de acordo com uma prioridade probabilisticamente definida. Nesse trabalho também foram usados os algoritmos *Random Forest* (BREIMAN, 2001) e NB-SVM (WANG; MANNING, 2012) que suportam classificação *multi-label*, treinando cada classe separadamente.

3 TRABALHOS RELACIONADOS

Sendo um dos primeiros trabalhos a estudar Usabilidade e Experiência do Usuário em postagens, Hedegaard & Simonsen (2013) investigam a presença de informações de diferentes dimensões de UUX em revisões online (textos escritos por usuários que usaram um produto por algum tempo, especificando prós e/ou contras dele, geralmente contendo uma avaliação ou recomendação para possíveis compradores). Inicialmente foi realizado um estudo de análise manual das sentenças, e em seguida, implementado e testado um classificador, visando primariamente conhecer o vocabulário relacionado aos aspectos de UUX considerados, e também analisar a viabilidade de aplicá-lo a uma grande base de dados. Nessa implementação, foi usado o algoritmo SVM, treinado com uma base de revisões representadas por Term Frequency & Inverse Document Frequency (TF-IDF) pré-processada com a biblioteca Natural Language Toolkit (NLTK). Apesar de aplicar um classificador bastante conhecido para uso em texto, os resultados apontam que ele tende a não ser tão eficiente, por causa do balanço de frequência das dimensões, e do complexo vocabulário associado a algumas delas. Como foi considerado um baixo volume de dados, não foi possível obter muitas informações de todas as facetas.

Os artigos de Freitas et al. (2016) e Silva et al. (2017) relatam a experiência de avaliar a UUX de sistemas usando a metodologia MALTU. Cada etapa é realizada manualmente, incluindo a classificação por facetas de UUX.

Mendes & Furtado (2017) apresentam a ferramenta UUX-Posts, com o intuito de auxiliar avaliações de UUX de sistemas a partir das PRUs. Para tentar automatizar os processos, é usado o Modelo Booleano (LARSON, 2010), que usa palavras-chave (padrões de identificação) para extrair informação, e os classificadores Árvore de Decisão e NB. A validação e treinamento dos algoritmos é feita com uma base de postagens classificada manualmente por especialistas. Os autores comparam a extração automática de padrões com uma estratégia manual baseada em hipóteses e análises estatísticas. É também citado a dificuldade da classificação manual e da automatização da extração e classificação.

Sumarizando, diferentemente deste trabalho, Silva et al. (2017) e Freitas et al. (2016) realizam classificação manual para avaliar Usabilidade e Experiência do Usuário em UUXs, mas estão relacionados também no sentido de entender como informações de UUX aparecem em diferentes contextos de postagens e de relatar a dificuldade do estudo manual. Mendes & Furtado (2017) e Hedegaard & Simonsen (2013) realizam classificação automática, porém com técnicas mais convencionais de aprendizado de máquina para o problema considerado.

4 METODOLOGIA

A metodologia dessa pesquisa é majoritariamente experimental, porém com um pouco de exploração, pois busca-se, ao longo do desenvolvimento de um modelo para classificação de postagens, também obter *insights* sobre como as informações de Usabilidade e Experiência do Usuário se apresentam em PRUs, de forma a guiar um melhor pré-processamento e consequentemente influenciar na classificação. O primeiro passo é obter bases de postagens classificadas manualmente por trabalhos anteriores. Nesses dados são realizadas análises básicas dos dados, procedimentos de pré-processamento de postagens e facetas, investigações sobre as bases de dados para um melhor entendimento do problema. Em seguida, são feitos testes com algoritmos de aprendizado para encontrar a melhor maneira de classificar facetas de UUX em PRUs.

As bases usadas para os experimentos são descritas na Tabela 1, onde ReclameAqui - Spotify (FREITAS *et al.*, 2016) foi extraída de um site de reclamações, PlayStore - Google Maps, PlayStore - Waze, WindowsStore - Waze (SILVA *et al.*, 2017) e PlayStore - MyFitnessPal foram extraídas de lojas de aplicativos, SIGAA (MENDES; FURTADO, 2017) foi extraída do Sistema Acadêmico usado na Universidade Federal do Ceará e Twitter (MENDES; FURTADO, 2017) da rede social.

Tabela 1 – Informações dos datasets, sendo N o tamanho, c o número de categorias identificadas nas postagens e T o total de facetas consideradas na classificação da base.

Dataset	N	c	T
ReclameAqui − Spotify	219	10	13
PlayStore − Google Maps	303	7	13
PlayStore − Waze	348	10	13
WindowsStore − Waze	362	8	13
SIGAA	649	13	13
Twitter	1938	13	22
PlayStore − MyFitnessPal	3581	12	13
Total	7405		

Fonte: O autor (2018).

A Figura 1 apresenta uma amostra dos dados brutos. A priori se percebe que é um problema de classificação multi-label, como explicado no capítulo de Fundamentação Teórica, sendo possível haver mais de uma faceta a cada PRU. Também é notável uma alta presença de

"Satisfação"e o uso de linguagem coloquial nos textos.

Figura 1 – Amostra das bases de dados.

	PRU	metaU	metaUX
1	Ótimo Não me deixou na mão ainda!!!	Satisfação	Satisfação
2	O melhor	Satisfação	Satisfação
3	O melhor	Satisfação	Satisfação
4	Bom D+ já me ajudou muito principalmente em lu...	Satisfação	Satisfação; Confiança
5	Está completamente errado!! Da caminhos e rota...	Eficácia	Frustração
6	É muito útil pra mim é esses dias começou com...	Utilidade	NaN
7	Meu mapa,minhas direções Ótimo msm (y) merece ...	Satisfação	Satisfação
8	É esse aplicativo é show de bola Antônio Jacin...	Satisfação	Satisfação
9	Tony Top eu uso muito eu faço entregas de lanc...	Eficácia; Eficiência; Utilidade; Satisfação	Satisfação; Confiança
10	Muito bom Sempre me ajuda	Utilidade; Satisfação	Satisfação

Fonte: O autor (2018).

4.1 Pré-processamento

Após uma análise prévia dos dados, para ter uma visão geral do que seria necessário aplicar à base, foi feito um pré-processamento desses *datasets*, com o intuito de aumentar a qualidade dos dados. Nesse processo, foram considerados as principais etapas para tratamento de dados: limpeza, integração e transformação dos dados (HAN *et al.*, 2011). Todas as bases foram unificadas em uma base de dados, denominada ao longo do texto como *"All together"*, e manipuladas pela biblioteca Pandas (MCKINNEY *et al.*, 2010).

Para as PRUs, foi realizada transformação para *case* minúscula, substituição de abreviaturas e gírias — comum ao contexto de onde os dados foram extraídos —, remoção de *stopwords* (palavras irrelevantes, como artigos e proposições), remoção de caracteres especiais (pontuação), stemização (reduzir uma palavra flexionada ao seu radical) e separação das sentenças em palavras (tokenização). A substituição de abreviaturas foi possível por meio de um pequeno dicionário minerado pelo autor [1]. Para a remoção de stopwords foi usada a biblioteca NLTK (BIRD; LOPER, 2004). Para a tarefa de stemização, foi usado o algoritmo Removedor de Sufixos da Lingua Portuguesa (RSLP) (HUYCK; ORENGO, 2001), também incluso na biblioteca NLTK. Um exemplo da aplicação dos métodos é mostrado no Quadro 5.

[1] Disponibilizado em https://gist.github.com/marcos1262/5c9dea0535072a30ab0ecc26b9fa3c1d

Quadro 5 – Exemplo da aplicação de pré-processamento.

Método	**Texto**
Original	Prático e eficiente! Mt bom!!!
Case minúscula	prático e eficiente! mt bom!!!
Substituição de abreviaturas e gírias	prático e eficiente! muito bom!!!
Remoção de *stopwords*	prático eficiente! bom!!!
Remoção de caracteres especiais	prático eficiente bom
Stemização	prát efici bom
Tokenização	["prát", "efici", "bom"]

Fonte: O autor (2018).

Para as facetas, foi aplicada transformação para *case* minúscula, remoção de espaços em branco, correção da grafia de algumas palavras e binarização dos rótulos, necessária para adaptar os dados para a forma de entrada dos algoritmos, como citado na seção de Fundamentação Teórica, e facilitar a análise das bases de dados.

Após ter as bases limpas e normalizadas, é necessário transformar os textos em *features* numéricas, para que a maquina possa aprender. As PRUs nessa etapa foram codificadas com TF-IDF para os algoritmos mais tradicionais e com *word embeddings* para os algoritmos de *Deep Learning*. Foram usados vetores de palavras próprios para a língua portuguesa (HARTMANN *et al.*, 2017).

A Figura 2 resume as etapas de pré-processamento usadas nos experimentos.

4.2 Análise dos dados

Com o objetivo de entender melhor os dados, de forma a descobrir informações úteis sobre o problema e ajudar na tomada de decisões durante os experimentos, as bases foram inspecionadas de diversas formas, quantidade de cada faceta nos *datasets*, relação entre as facetas, tamanho das PRUs e quantidade de palavras únicas. Para essa tarefa foi usada a biblioteca Seaborn (WASKOM *et al.*, 2018), baseada na renomada biblioteca Matplotlib (HUNTER, 2007).

4.3 Classificação

Para os classificadores de Machine Learning foi usada a biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011). Para as abordagens de transformação de problema e adaptação de algoritmos para multi-label foi usada a biblioteca *Scikit-Multilearn* (Szymański; Kajdanowicz, 2017). Para os modelos de Deep Learning foi usada a biblioteca *Keras* (CHOLLET *et al.*, 2015).

Figura 2 – Etapas de pré-processamento para features e labels.

(a) Pré-processamento das Postagens.

(b) Pré-processamento das Facetas.

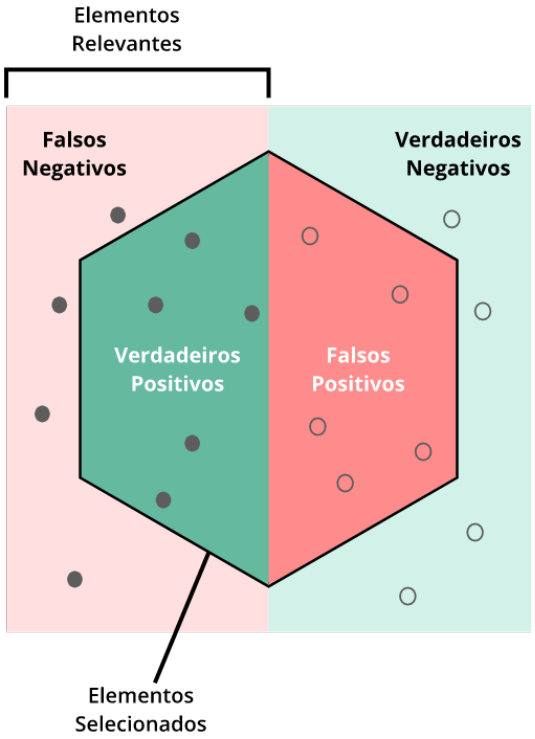
Fonte: O autor (2018).

4.3.1 Métricas

Foram consideradas as principais métricas para classificação binária: acurácia, precisão, *recall* e F-measure (F1). Elas são baseadas nas quantidades de erros e acertos do classificador a ser avaliado, ilustradas na Figura 3. O losango no centro refere-se às predições positivas feitas pelo algoritmo. O lado esquerdo da figura representa os elementos que o classificador deveria ter escolhido como positivo. Com essas informações podemos calcular, por exemplo, o número de falsos positivos, que são os elementos que deveriam ser classificados negativamente mas foram classificados falsamente como positivos.

As métricas usadas são calculadas com bases nesses acertos e erros de acordo com as equações 4.1, 4.2, 4.3 e 4.4. A acurácia significa a porcentagem de items que foram acertados no total, a precisão expressa quantos itens selecionados são relevantes (quantos itens o classificador deveria acertar, de todos os que ele selecionou como positivo), o *recall* representa a porcentagem de elementos relevantes que foram selecionados e a métrica F1 calcula uma média harmônica

Figura 3 – Acertos e erros de um classificador.



Fonte: O autor (2018).

entre precisão e recall.

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN}. \tag{4.1}$$

$$Precisão = \frac{VP}{VP+FP}. \tag{4.2}$$

$$Recall = \frac{VP}{VP+FN}. \tag{4.3}$$

$$F1 = 2 \cdot \frac{precisão \cdot recall}{precisão + recall}. \tag{4.4}$$

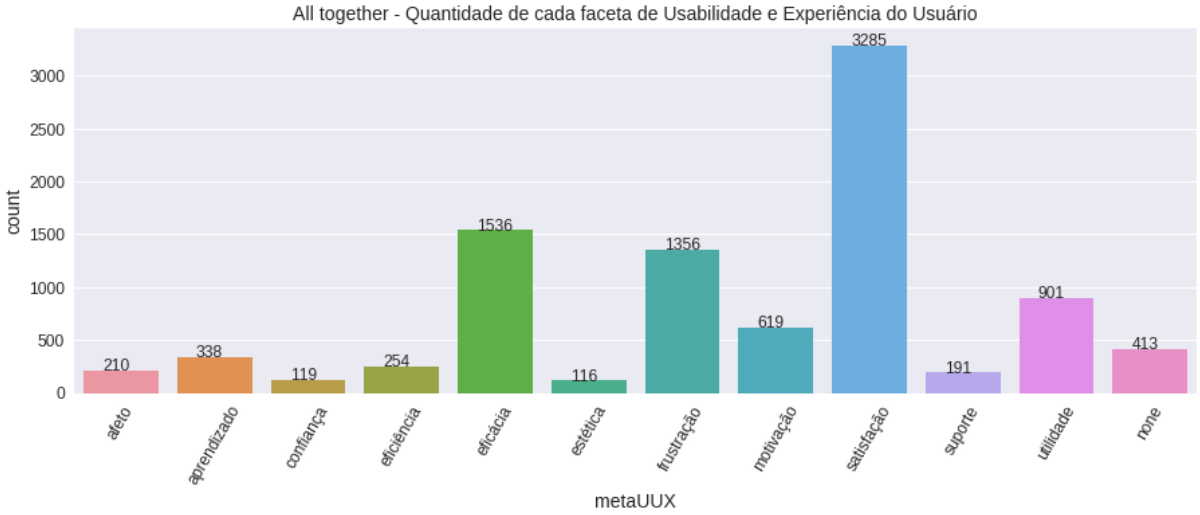
Estas porém, na sua forma padrão, foram projetadas para classificação *single-label*, onde é associada somente uma classe para cada item no *dataset*, e por isso devem ser adaptados. Neste trabalho as métricas foram implementadas como a média de suas pontuações para cada classe.

4.3.2 Experimentos

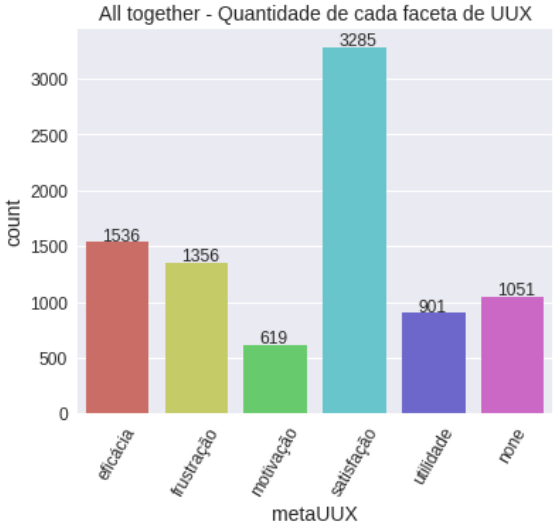
Os experimentos foram realizados em uma plataforma na nuvem. O *dataset* foi dividido em 70% para treino e 30% para teste. As pontuações definitivas foram calculadas como a média de 10 execuções embaralhando os dados. Em cada iteração, foi usada exatamente a mesma base para todos os algoritmos.

Também foram realizados experimentos excluindo algumas facetas a fim de tentar melhorar a base de dados. Primeiramente, foram removidas as facetas que tinham quantidade de exemplos menor que 1% da quantidade total (7405). Em seguida, foram realizados testes removendo facetas com quantidades menores que 5%, 10% e 20%. Às PRUs com nenhuma faceta relacionada foi atribuído o rótulo *"none"*. A Figura 4 mostra a distribuição de PRUs por rótulo após cada filtragem.

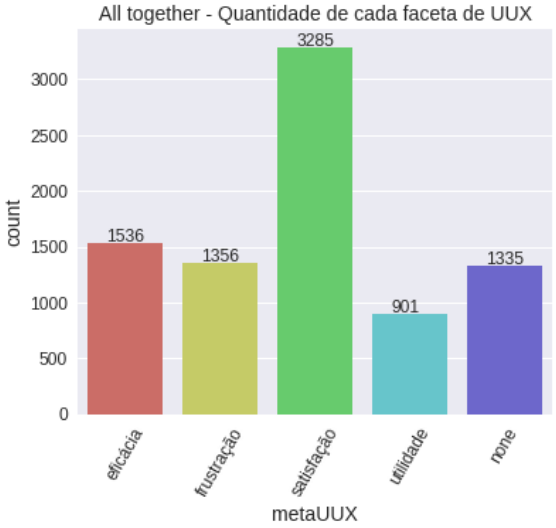
Figura 4 – Quantidade de cada faceta de UUX nas bases de dados com número reduzido de facetas.



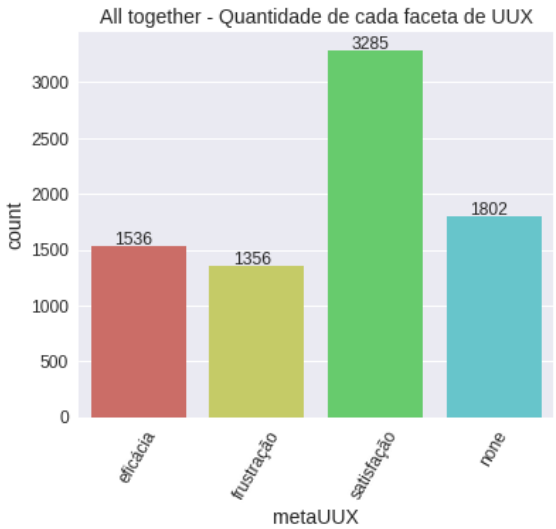
(a) Menos 1%.



(b) Menos 5%.



(c) Menos 10%.



(d) Menos 20%.

Fonte: O autor (2018).

5 RESULTADOS

Nessa seção estão contidos os resultados obtidos com os experimentos e classificações das PRUs em relação às facetas de UUX. São apresentadas algumas percepções expressas em gráficos, que sumarizam as informações pesquisadas nos dados, e as pontuações atingidas pelos algoritmos para cada métrica considerada.

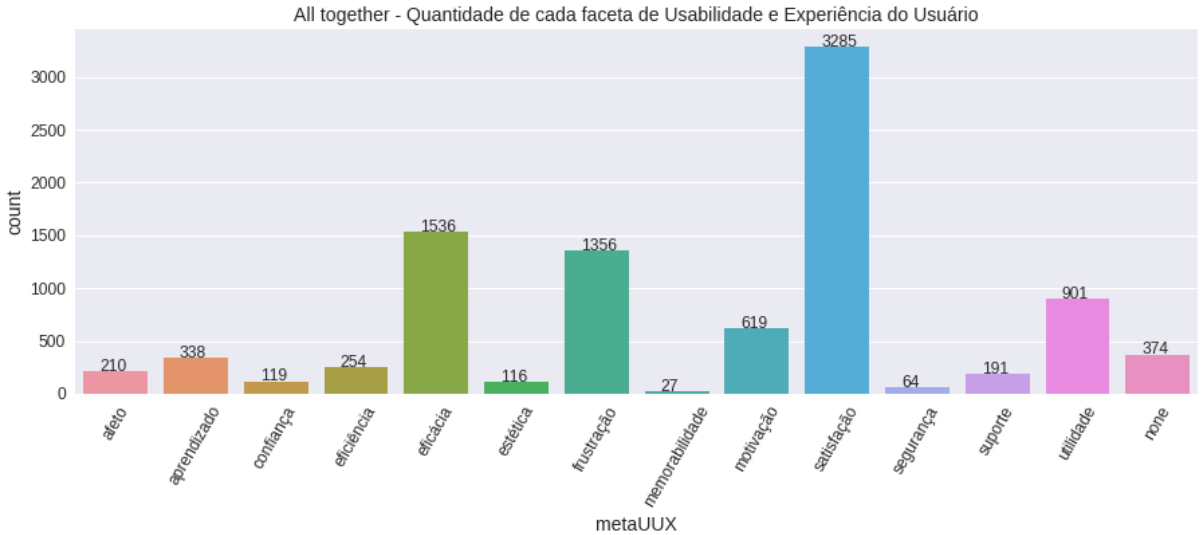
Alguns dos gráficos mostrados a seguir referem-se ao conjunto completo de PRUs analisadas, agregadas na base de dados *"All together"*. As plotagens para cada base individual se encontram no Apêndice A.

5.1 Análise dos Dados

Os gráficos gerados na análise das bases de dados permitiram tomar as decisões em relação aos experimentos de classificação e ter um melhor conhecimento do problema de apontar as facetas relacionadas a uma Postagem Relacionada ao Uso.

A Figura 5 mostra a quantidade de cada meta de Usabilidade e Experiência do Usuário presente no conjunto completo de dados. Também é contabilizado o pequeno número de PRUs que não possuem nenhuma faceta relacionada, representado pelo rótulo *"none"*.

Figura 5 – Quantidade de cada faceta de UUX.



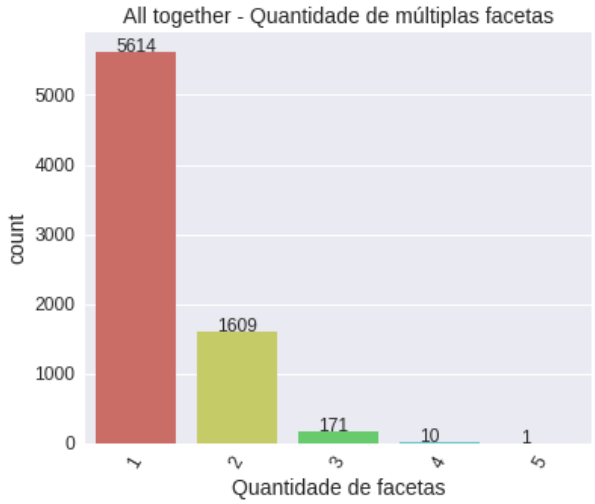
Fonte: O autor (2018).

Nota-se uma diferença discrepante entre as facetas, tendo o máximo 3285 ("Satisfação") e o mínimo 27 ("Memorabilidade"). Em um balanceamento perfeito da base, cada faceta

teria por volta de 570 exemplos, porém na realidade a maioria não chega perto desse valor, enquanto outras ultrapassam demasiadamente.

Por se estar lidando com um problema *multi-label*, o somatório das quantidades no gráfico resulta em um valor maior que a quantidade de PRUs, dado que algumas delas estão relacionadas com múltiplas facetas. O número de múltiplos rótulos é mostrado na Figura 6, onde pode-se destacar que à maioria das PRUs foi atribuída somente uma faceta, e que há uma outra parcela significante com duas metas.

Figura 6 – Quantidade de PRUs com múltiplas facetas.



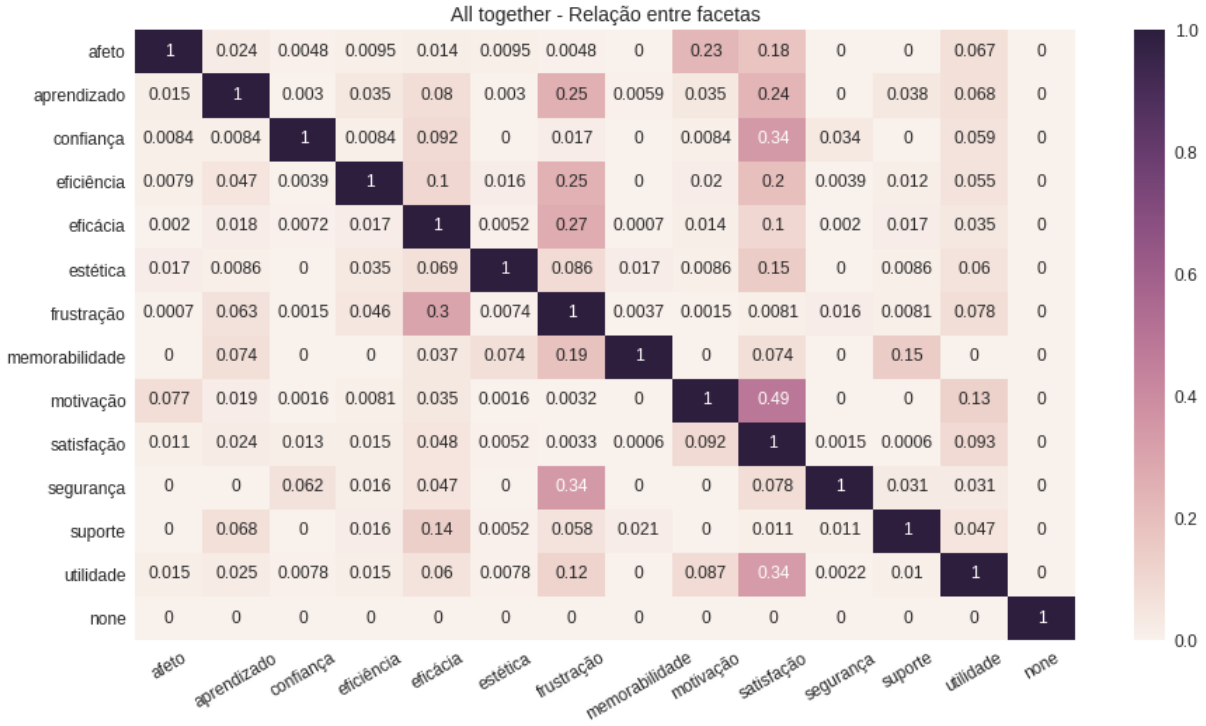
Fonte: O autor (2018).

A Figura 7 exprime a relação entre cada faceta, analisando a porcentagem de aparições conjuntas delas. Por exemplo, 23% das PRUs relacionadas a "Afeto"também aparecem como "Motivação", como pode ser observado na primeira linha do gráfico. Mesmo com a maioria das facetas aparecendo sozinha (PRUs com somente uma faceta), pode-se ver alguns resultados significativos como 49% das PRUs sobre "Motivação"também revelarem "Satisfação", indicando que frequentemente usuários que se sentem motivados no sistema também demonstram satisfação, e 30% das PRUs sobre "Frustração"serem relacionadas à "Eficácia"do sistema.

Considerando os datasets independentemente (Apêndice A), nota-se números ainda mais expressivos. Nas análises de cada base de dados, essas informações são muito valiosas para entender e poder avaliar a Usabilidade e Experiência do Usuário do sistema em questão.

Como visto na descrição dos *datasets* na Tabela 1, alguns não contêm todas as facetas consideradas nessa pesquisa. Essa distribuição pode ser melhor analisada para cada faceta na Figura 8, que apresenta a quantidade de bases que englobam cada meta de UUX. Do lado direito

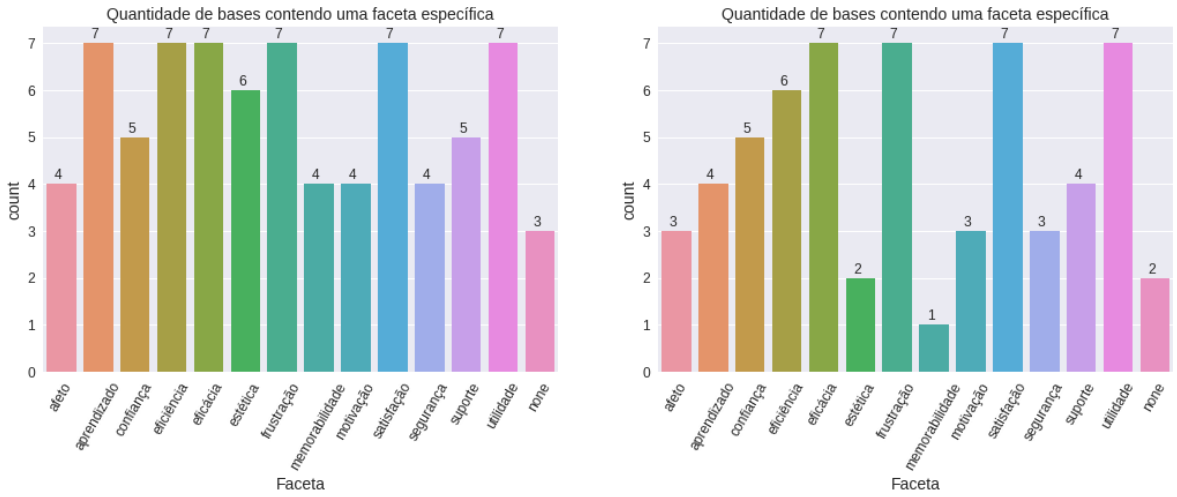
Figura 7 – Relação entre facetas.



Fonte: O autor (2018).

está a plotagem desconsiderando números minúsculos (facetas que apareciam somente em até 7 PRUs – menos que $0,01\%$ da quantidade total).

Figura 8 – Presença de cada faceta nas bases de dados.



(a) Dados completos (b) Dados mais significativos

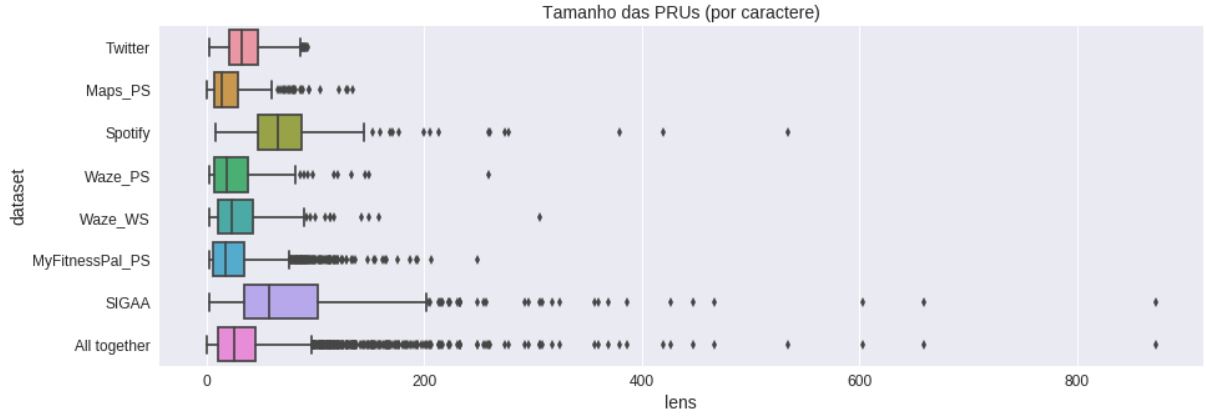
Fonte: O autor (2018).

A partir dos gráficos exibidos acima, pode-se ter uma noção sobre o balanceamento da base de PRUs. Algumas facetas aparecem em quantidades bastante diferentes enquanto outras

têm uma representatividade muito baixa diante de todo o conjunto de dados. Essas características podem influenciar muito nos resultados gerais de um classificador, visto que, por exemplo, ele provavelmente terá maior facilidade para identificar determinadas facetas nas PRUs do que outras, o que posteriormente levará a uma pontuação não tão satisfatória.

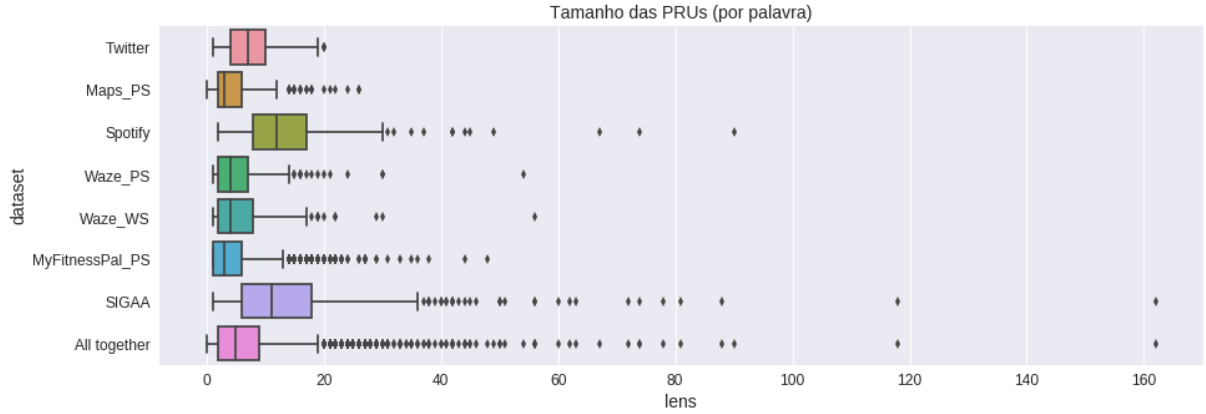
Os gráficos apresentados nas Figuras 9, 10 e 11, foram importantes para definir alguns parâmetros para os experimentos, como o tamanho de cada postagem e a quantidade de palavras únicas para se poder otimizar a conversão das PRUs para features numéricas. As PRUs foram cortadas em um único tamanho, que foi definido como o valor que descartasse tamanhos excessivos perdendo uma menor parte possível dos textos. No caso da separação por palavras (resultado do processo de tokenização explicado na seção de Metodologia), as postagens foram limitadas a um tamanho por volta de 60 palavras.

Figura 9 – Tamanho das PRUs (por caractere).



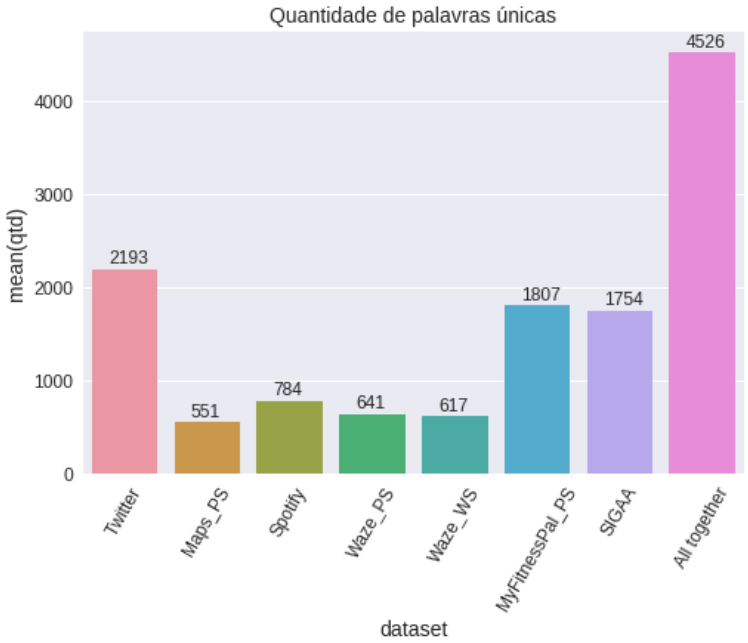
Fonte: O autor (2018).

Figura 10 – Tamanho das PRUs (por palavra).



Fonte: O autor (2018).

Figura 11 – Quantidade de palavras únicas.



Fonte: O autor (2018).

Explorações a nível de *dataset* podem indicar os sentimentos expressos pelos usuários relacionados ao sistema em que foram publicadas as PRUs. Análises semelhantes a da Figura 5 e da Figura 7 também são muito importantes ao classificar, seja manualmente ou automaticamente, PRUs de um sistema para obter conclusões sobre a opinião dos usuários sobre o sistema.

5.2 Classificação

O resultado geral é apresentado na Tabela 2, onde estão destacados os valores mais altos sobre todos e também os mais altos dentre os algoritmos de Deep Learning. O classificador SVM obteve os melhores resultados, variando um pouco sua pontuação entre os métodos de transformação do problema multi-label. Dentre os modelos de Deep Learning aplicados, o algoritmo com GRU obteve os melhores resultados.

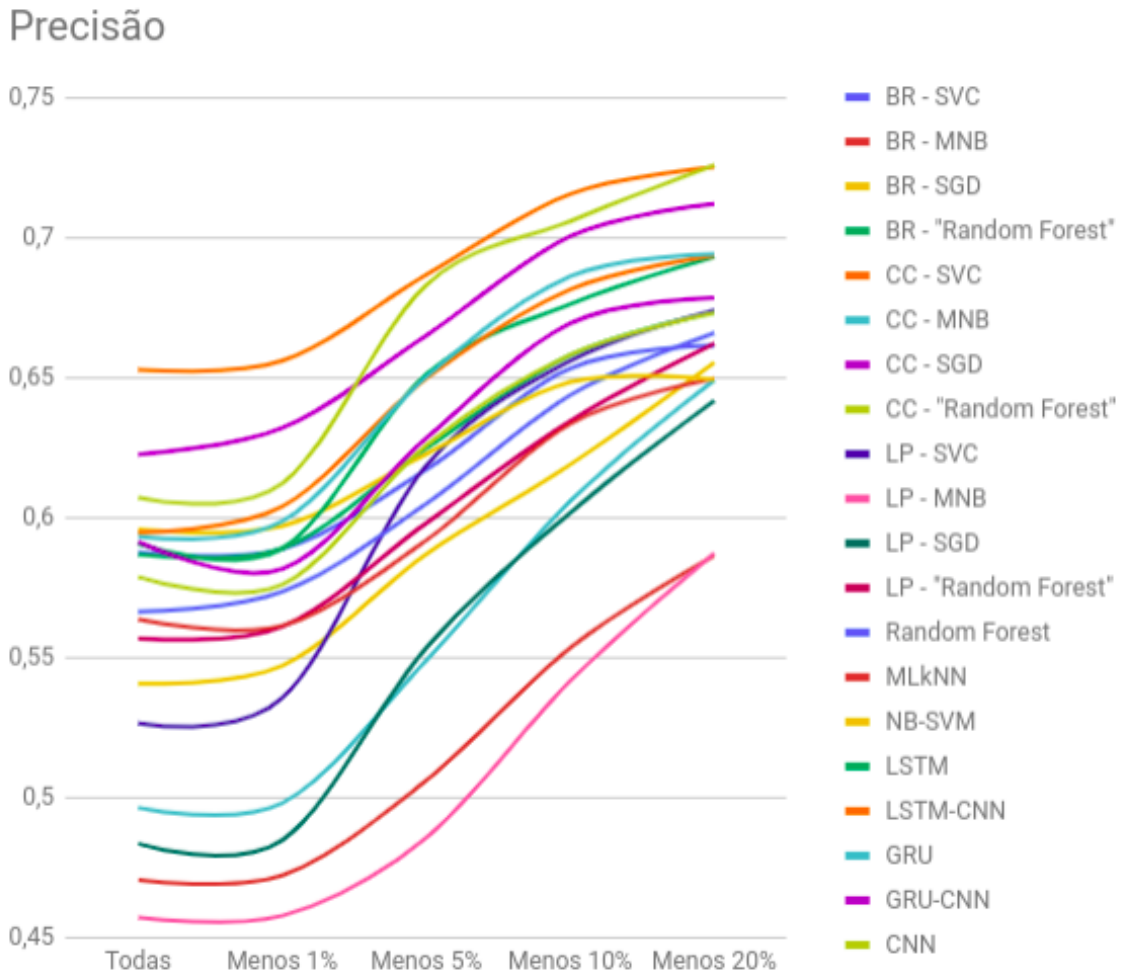
A partir dos experimentos com número reduzido de facetas, pôde-se analisar a melhoria das pontuações, expressa pelos gráficos da Figura 12, que mostram também as linhas de tendência linear de cada algoritmo, destacando os mais inclinados. Na Figura 12(c) se destacam Random Forest com Classifier Chain, SVM e SGD com Label Powerset, respectivamente de cima para baixo.

Tabela 2 – Pontuações de cada algoritmo de aprendizado, destacando as maiores gerais e as maiores de Deep Learning.

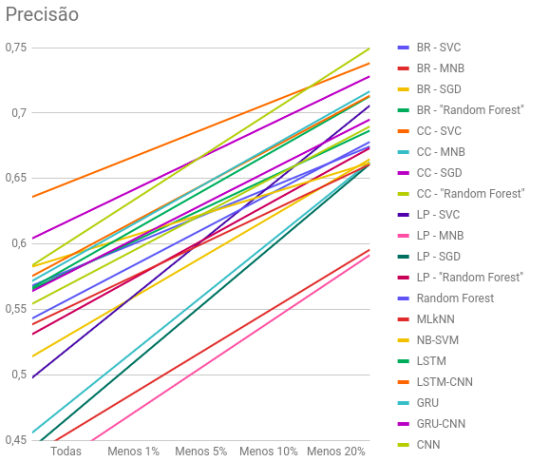
Algoritmo	**Acurácia**	**Precisão**	**Recall**	**F1**	**Média**
Binary Relevance − SVM	**0,94735**	0,5874	**0,74513**	0,594	0,71847
Binary Relevance − Naive Bayes	0,94225	0,47027	0,72543	0,48922	0,6567925
Binary Relevance − SGD	0,94536	0,54049	0,73699	0,55153	0,6935925
Binary Relevance − Random Forest	0,94195	0,58669	0,71214	0,58295	0,7059325
Classifier Chain − SVM	0,93883	**0,65266**	0,68934	**0,65553**	**0,73409**
Classifier Chain − Naive Bayes	0,94301	0,49617	0,72894	0,51587	0,6709975
Classifier Chain − SGD	0,93677	0,62241	0,67411	0,63406	0,7168375
Classifier Chain − Random Forest	0,94032	0,60687	0,7024	0,60456	0,7135375
Label Powerset − SVM	0,94523	0,52636	0,74043	0,54179	0,6884525
Label Powerset − Naive Bayes	0,94066	0,45695	0,71737	0,47779	0,6481925
Label Powerset − SGD	0,94329	0,48329	0,73016	0,49846	0,6638
Label Powerset − Random Forest	0,9404	0,55663	0,70442	0,56445	0,691475
Adapted Random Forest	0,94052	0,56614	0,70418	0,57349	0,6960825
MLkNN	0,94155	0,56345	0,71404	0,57132	0,69759
NB−SVM	0,94571	0,59556	0,73449	0,59632	0,71802
LSTM	0,9436	0,59047	0,71964	0,59576	0,7123675
LSTM−CNN	0,9426	**0,59446**	0,7139	**0,59888**	0,71246
GRU	**0,94519**	0,593	**0,72909**	0,59516	**0,71561**
GRU−CNN	0,94372	0,59114	0,72082	0,5902	0,71147
CNN	0,9449	0,57858	0,72798	0,58116	0,708155

Fonte: O autor (2018).

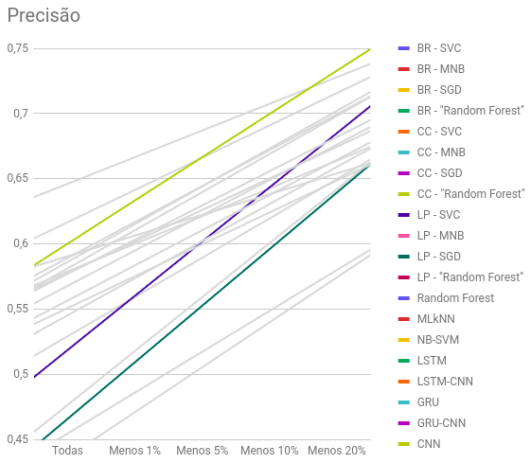
Figura 12 – Precisão de cada algoritmo reduzindo a quantidade de facetas.



(a) Crescimento das pontuações.



(b) Tendência linear.



(c) Possivelmente melhores destacados.

Fonte: O autor (2018).

6 CONCLUSÕES E TRABALHOS FUTUROS

6.1 Considerações Finais

Esse estudo buscou desenvolver um modelo para classificação automática de Postagens Relacionadas ao Uso segundo as metas de Usabilidade e Experiência do Usuário, consistindo na preparação dos textos e rótulos, investigação sobre os dados obtidos para treinamento, escolha de algoritmos para realizar a classificação.

A tarefa de classificação textual em questão, denotada como um problema de classificação *multi-label*, foi abordada por métodos conhecidos na literatura para tratar múltiplos rótulos, em conjunto com técnicas mais tradicionais de *Machine Learning*, incluindo o algoritmo de aprendizado *Support Vector Machines*, que tem sido estado-da-arte nessa área específica e vem sendo superado nos anos recentes.

Em geral, os métodos que consideram alguma relação entre as facetas (*Classifier Chain* e *Label Powerset*) obtiveram melhores resultados que o de relevância binária (que trata cada classe independentemente), o que indica que a relação entre as facetas tenha alguma importância na classificação.

Nos experimentos realizados, SVM apresentou os melhores resultados, seguido de SGD, GRU, LSTM com CNN e *Random Forest*. Embora os algoritmos de *Deep Learning* não tenham obtido os melhores resultados, eles atingiram pontuações bem próximas às do algoritmo SVM e se mostram como uma promessa de estado-da-arte para a classificação de PRUs em relação às metas de UUX, possivelmente a partir do crescimento do acervo de postagens ou de um melhor balanceamento e distribuição de exemplos classificados.

6.2 Trabalhos Futuros

Como trabalhos futuros, podem ser consideradas mais abordagens para tratar o balanceamento dos *datasets*, ser investigadas mais possibilidades de algoritmos e de modelos de *Deep Learning*, e realizar uma otimização de parâmetros mais minuciosa. Uma classificação por cada faceta individualmente também pode ser testada, de forma a analisar o quanto cada faceta impactou no resultado geral apresentado por este estudo. Podem ser feitas mais análises como das palavras mais frequentes e/ou relevantes para cada faceta e experimentadas mais possibilidades de pré-processamento. Uma interface para facilmente realizar classificações pode

ser construída e também integrada ao sistema UUX-Posts para análise e classificação de PRUs.

REFERÊNCIAS

BARGAS-AVILA, J. A.; HORNBÆK, K. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: ACM. **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. [S.l.], 2011. p. 2689–2698.

BEVAN, N. Classifying and selecting ux and usability measures. In: **International Workshop on Meaningful Measures: Valid Useful User Experience Measurement**. [S.l.: s.n.], 2008. v. 11, p. 13–18.

BIRD, S.; LOPER, E. Nltk: the natural language toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL 2004 on Interactive poster and demonstration sessions**. [S.l.], 2004. p. 31.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

CHOLLET, F. *et al.* **Keras**. 2015. Disponível em: <https://keras.io>. Acesso em: 20 out. 2018.

CONNEAU, A.; SCHWENK, H.; BARRAULT, L.; LECUN, Y. Very deep convolutional networks for text classification. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**. [S.l.: s.n.], 2017. v. 1, p. 1107–1116.

FREITAS, L.; SILVA, T.; MENDES, M. Avaliação do spotify–uma experiência de avaliação textual utilizando a metodologia maltu. In: **Proceedings of the 15th Brazilian Symposium on Human Factors in Computer Systems (IHC'16)**. [S.l.: s.n.], 2016. v. 50.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v. 1.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. **arXiv preprint arXiv:1708.06025**, 2017.

HEDEGAARD, S.; SIMONSEN, J. G. Extracting usability and user experience information from online user reviews. In: ACM. **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. [S.l.], 2013. p. 2089–2098.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing In Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

HUYCK, C.; ORENGO, V. A stemming algorithmm for the portuguese language. In: **String Processing and Information Retrieval, International Symposium on(SPIRE)**. [S.l.: s.n.], 2001. v. 00, p. 0186.

ISO. **ISO 9241-210:2010 - Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems**. 2010. Disponível em: <http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=52075>. Acesso em: 10 jun. 2018.

KETOLA, P.; ROTO, V. Exploring user experience measurement needs. In: **Proc. of the 5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement (VUUM). Reykjavik, Island**. [S.l.: s.n.], 2008. p. 23–26.

KIM, Y. Convolutional neural networks for sentence classification. **arXiv preprint arXiv:1408.5882**, 2014.

KORHONEN, H.; ARRASVUORI, J.; VÄÄNÄNEN-VAINIO-MATTILA, K. Let users tell the story: evaluating user experience with experience reports. In: ACM. **CHI'10 Extended Abstracts on Human Factors in Computing Systems**. [S.l.], 2010. p. 4051–4056.

LAI, S.; XU, L.; LIU, K.; ZHAO, J. Recurrent convolutional neural networks for text classification. In: **AAAI**. [S.l.: s.n.], 2015. v. 333, p. 2267–2273.

LARSON, R. R. Introduction to information retrieval. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 61, n. 4, p. 852–853, 2010.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015.

LIMA, A. M.; SILVA, P. B.; CRUZ, L. A.; MENDES, M. S. Investigating the polarity of user postings in a social system. In: SPRINGER. **International Conference on Social Computing and Social Media**. [S.l.], 2017. p. 246–257.

MCKINNEY, W. *et al.* Data structures for statistical computing in python. In: AUSTIN, TX. **Proceedings of the 9th Python in Science Conference**. [S.l.], 2010. v. 445, p. 51–56.

MENDES, M. **MALTU-Model for evaluation of interaction in social systems from the Users Textual Language. 200 f**. Tese (Doutorado) — Federal University of Ceará (UFC), Fortaleza, CE–Brazil, 2015.

MENDES, M. S.; FURTADO, E.; FURTADO, V.; CASTRO, M. F. de. Investigating usability and user experience from the user postings in social systems. In: SPRINGER. **International Conference on Social Computing and Social Media**. [S.l.], 2015. p. 216–228.

MENDES, M. S.; FURTADO, E. S. Uux-posts: a tool for extracting and classifying postings related to the use of a system. In: ACM. **Proceedings of the 8th Latin American Conference on Human-Computer Interaction**. [S.l.], 2017.

MENDES, M. S.; FURTADO, E. S.; CASTRO, M. F. de. Do users write about the system in use?: an investigation from messages in natural language on twitter. In: ACM. **Proceedings of the 7th Euro American Conference on Telematics and Information Systems**. [S.l.], 2014. p. 3.

MENDES, M. S.; FURTADO, E. S.; THEOPHILO, F.; FRANKLIN, M. A study about the usability evaluation of social systems from messages in natural language. In: **Human Computer Interaction**. [S.l.]: Springer, 2013. p. 59–62.

OLSSON, T.; SALO, M. Narratives of satisfying and unsatisfying experiences of current mobile augmented reality applications. In: ACM. **Proceedings of the SIGCHI conference on human factors in computing systems**. [S.l.], 2012. p. 2779–2788.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

READ, J.; PFAHRINGER, B.; HOLMES, G.; FRANK, E. Classifier chains for multi-label classification. **Machine learning**, Springer, v. 85, n. 3, p. 333, 2011.

ROGERS, Y.; SHARP, H.; PREECE, J. **Design de interação: além da interação humano-computador**. [S.l.]: Bookman, 2013.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Malaysia; Pearson Education Limited,, 2016.

SILVA, T. H. O. da; FREITAS, L. M.; MENDES, M. S. Beyond traditional evaluations: user's view in app stores. In: ACM. **Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems**. [S.l.], 2017. p. 15.

Szymański, P.; Kajdanowicz, T. A scikit-based Python environment for performing multi-label classification. **ArXiv e-prints**, fev. 2017.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining (IJDWM)**, IGI Global, v. 3, n. 3, p. 1–13, 2007.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Mining multi-label data. In: **Data mining and knowledge discovery handbook**. [S.l.]: Springer, 2009. p. 667–685.

WANG, S.; MANNING, C. D. Baselines and bigrams: Simple, good sentiment and topic classification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2**. [S.l.], 2012. p. 90–94.

WASKOM, M.; BOTVINNIK, O.; O'KANE, D.; HOBSON, P.; OSTBLOM, J.; LUKAUSKAS, S.; GEMPERLINE, D. C.; AUGSPURGER, T.; HALCHENKO, Y.; COLE, J. B.; WARMENHOVEN, J.; RUITER, J. de; PYE, C.; HOYER, S.; VANDERPLAS, J.; VILLALBA, S.; KUNTER, G.; QUINTERO, E.; BACHANT, P.; MARTIN, M.; MEYER, K.; MILES, A.; RAM, Y.; BRUNNER, T.; YARKONI, T.; WILLIAMS, M. L.; EVANS, C.; FITZGERALD, C.; BRIAN; QALIEH, A. **mwaskom/seaborn: v0.9.0 (July 2018)**. 2018. Disponível em: <https://doi.org/10.5281/zenodo.1313201>. Acesso em: 20 out. 2018.

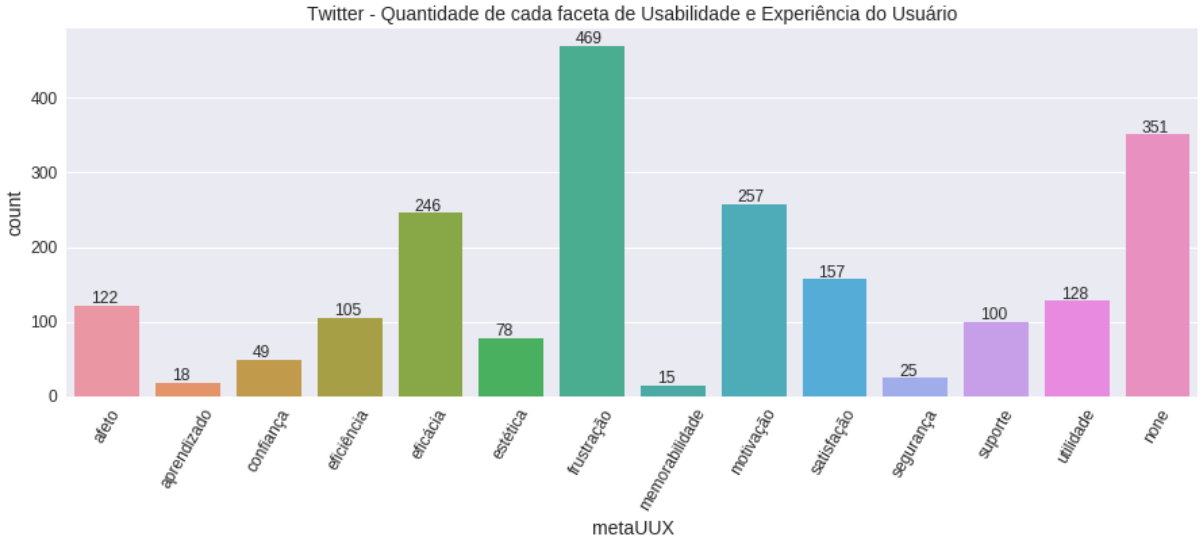
ZHANG, M.-L.; LI, Y.-K.; LIU, X.-Y.; GENG, X. Binary relevance for multi-label learning: An overview. **Front. Comput. Sci.**, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 12, n. 2, p. 191–202, abr. 2018. ISSN 2095-2228.

ZHANG, M.-L.; ZHOU, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. **Pattern recognition**, Elsevier, v. 40, n. 7, p. 2038–2048, 2007.

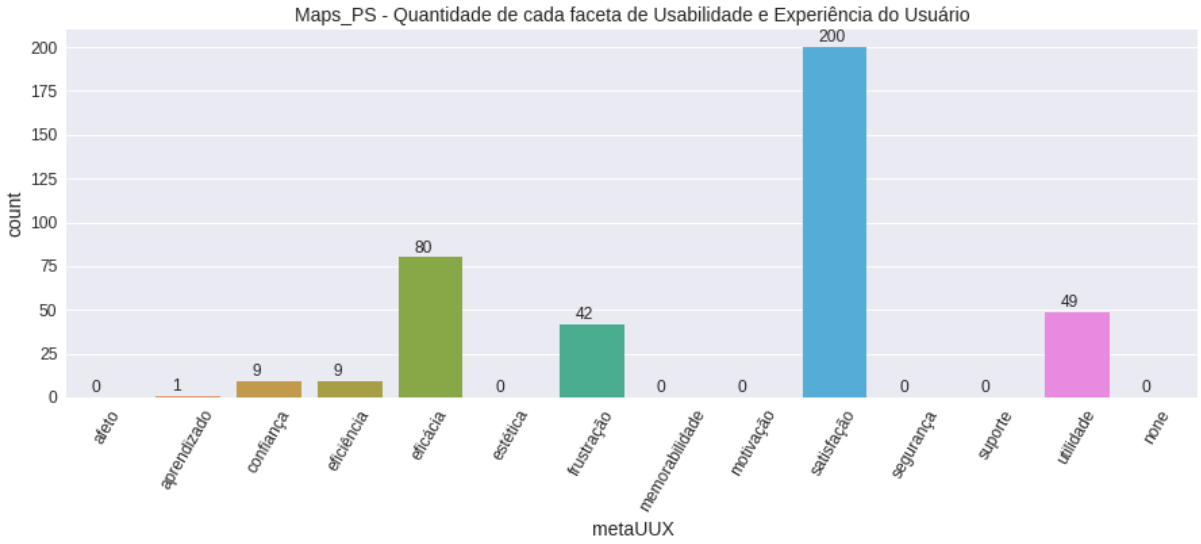
ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2015. p. 649–657.

APÊNDICE A – GRÁFICOS PARA CADA *DATASET*

Figura 13 – Quantidade de cada faceta de UUX para cada *dataset*.

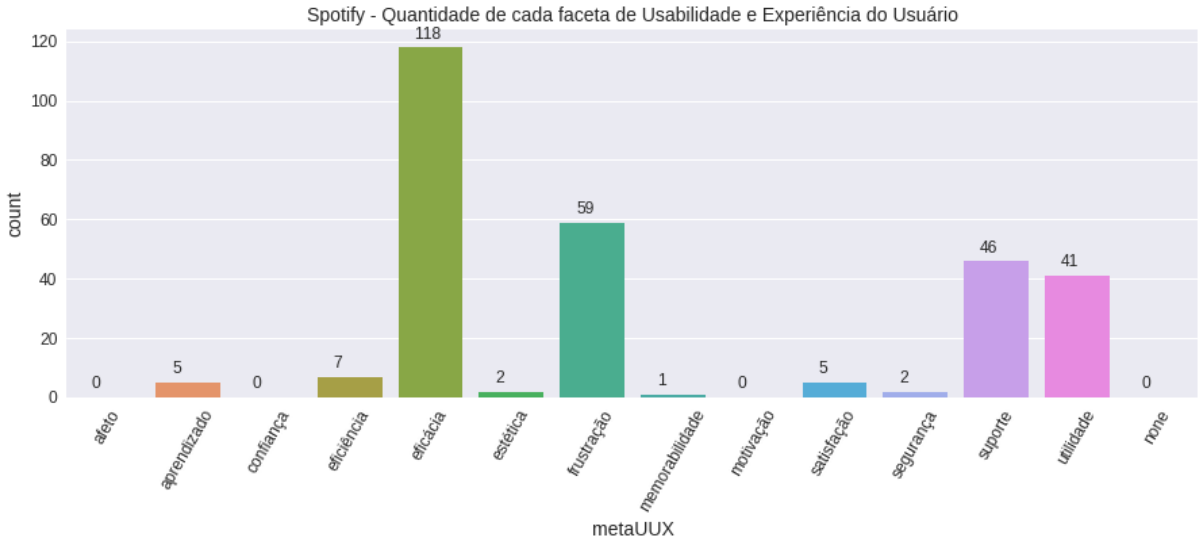


(a) Twitter

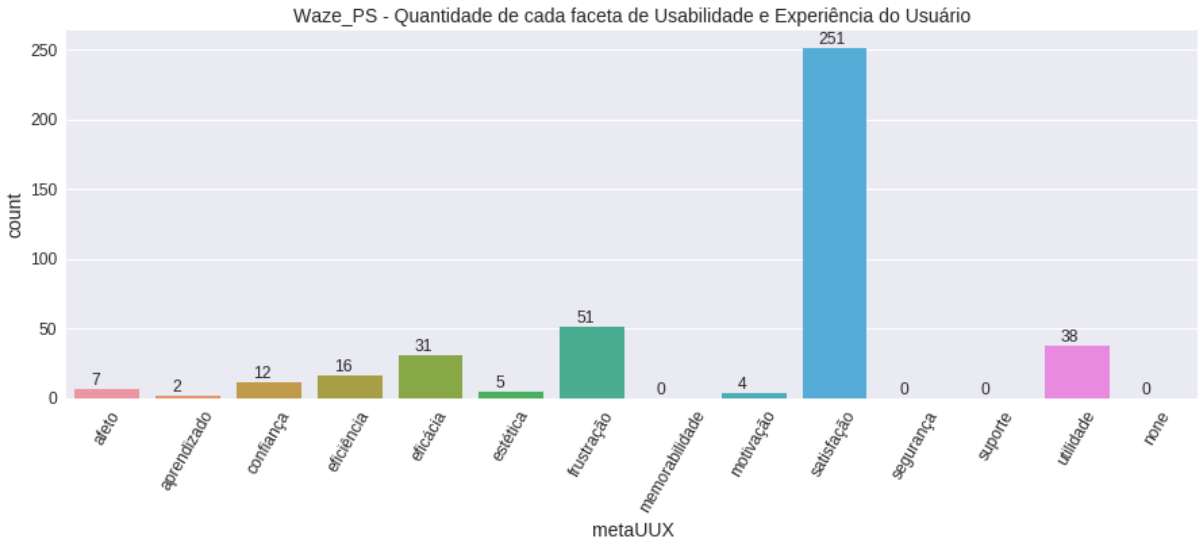


(b) PlayStore – Google Maps

Figura 13 – (continuação) Quantidade de cada faceta de UUX para cada *dataset*.

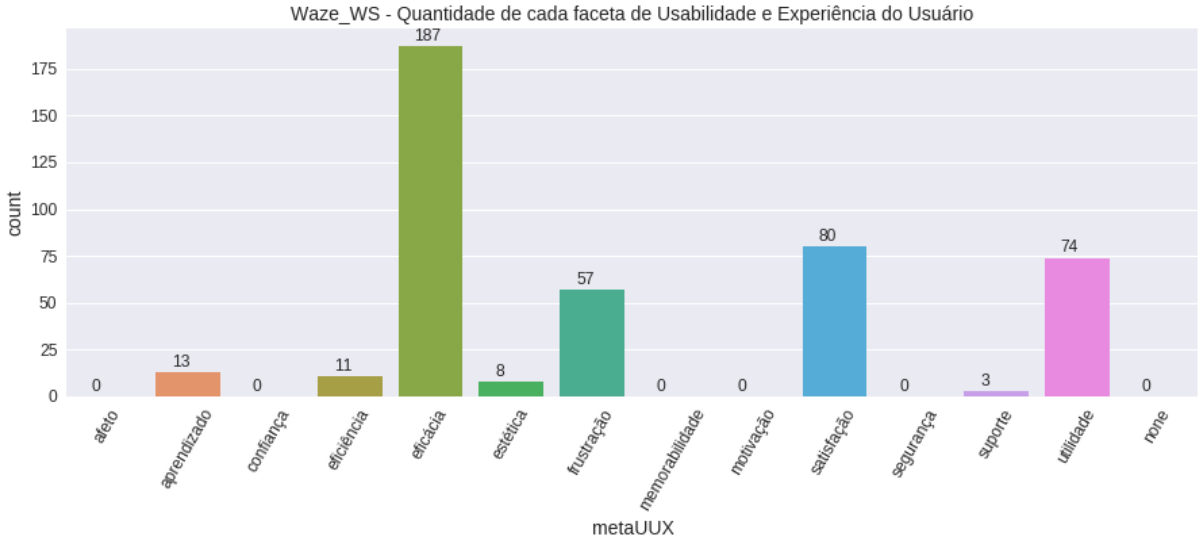


(a) ReclameAqui – Spotify

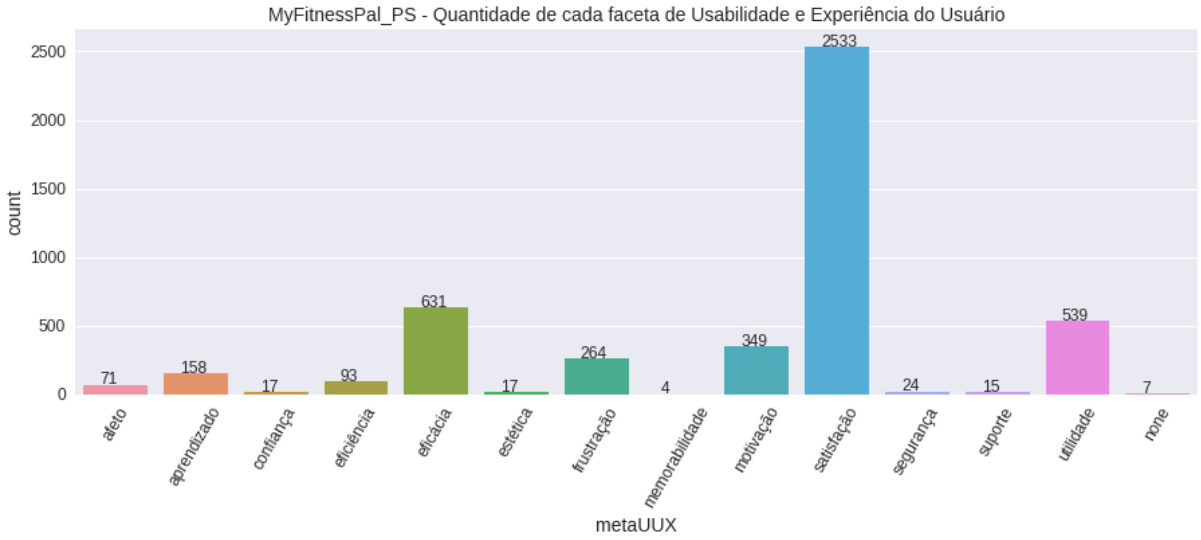


(b) PlayStore – Waze

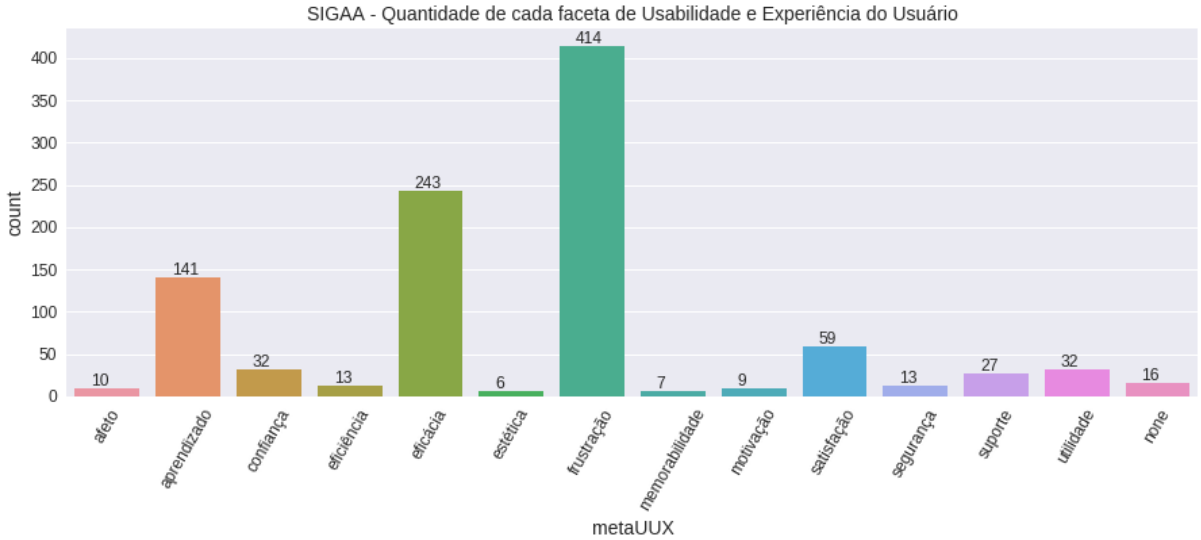
Figura 13 – (continuação) Quantidade de cada faceta de UUX para cada *dataset*.



(a) WindowsStore – Waze



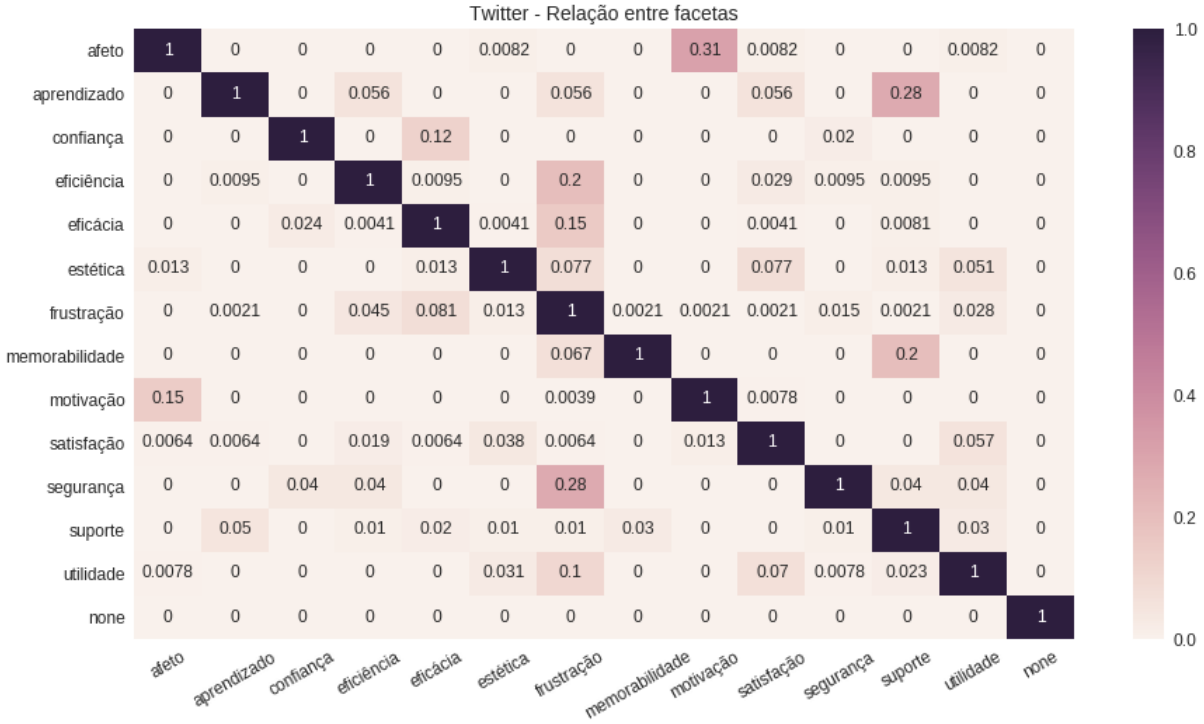
(b) PlayStore – MyFitnessPal



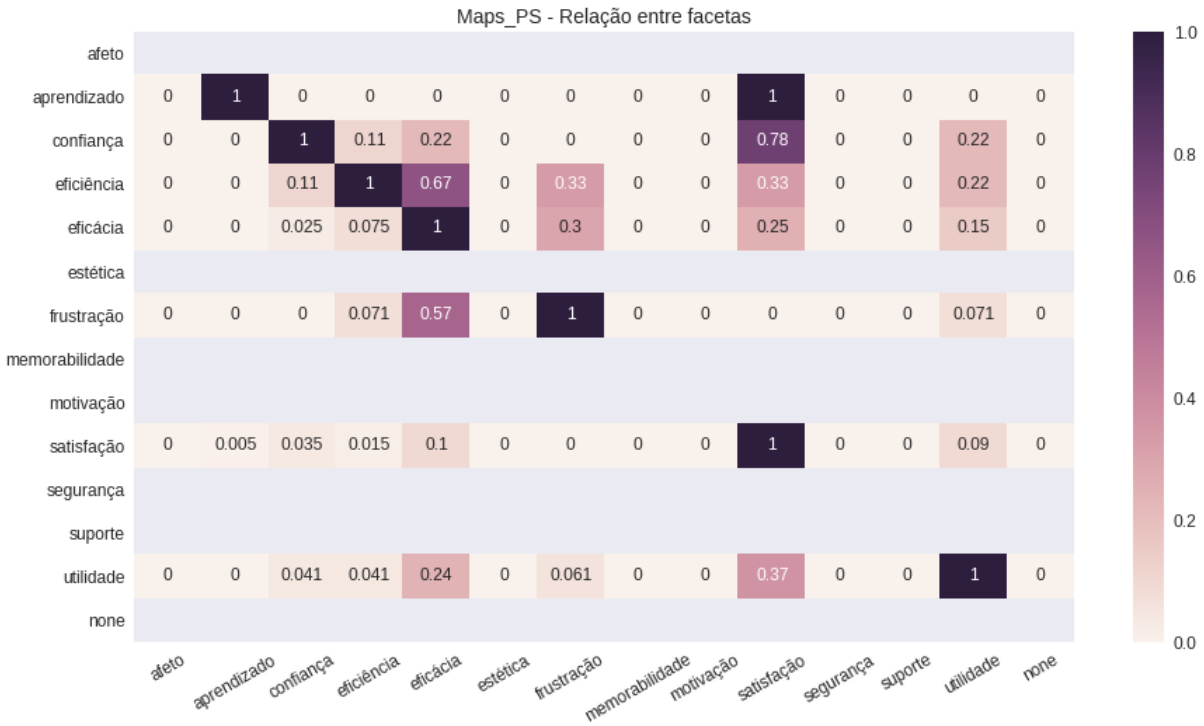
(c) SIGAA

Fonte: O autor (2018).

Figura 14 – Relação entre facetas para cada *dataset*.

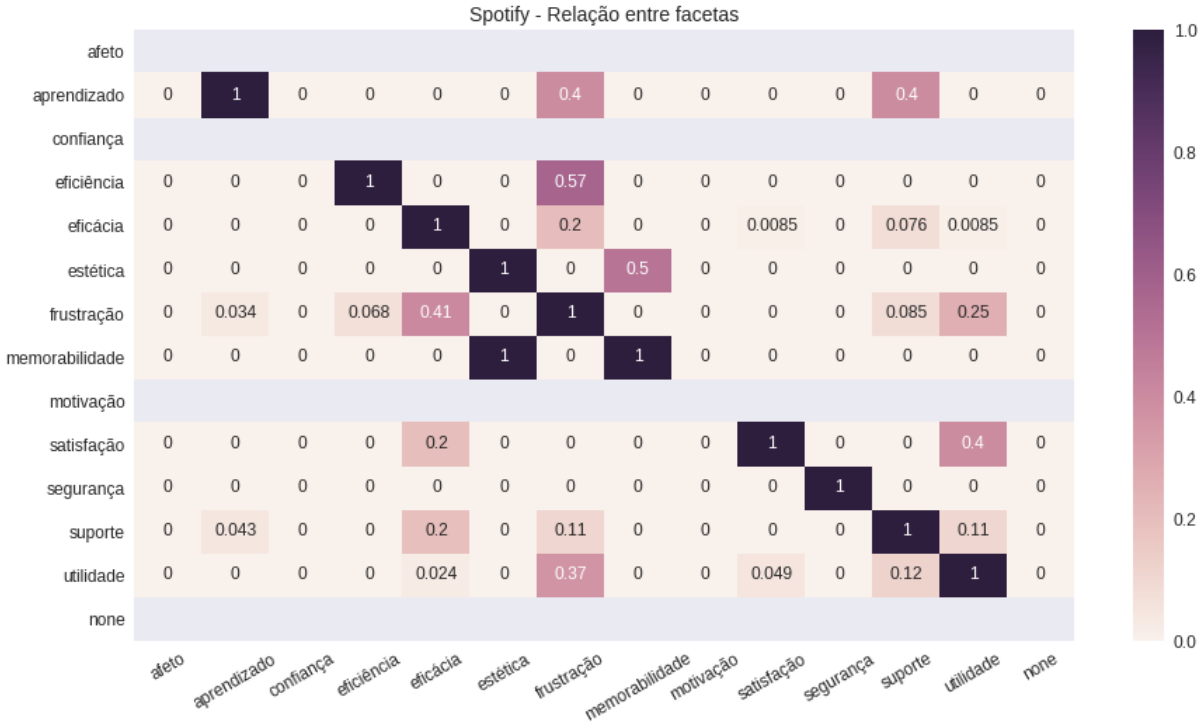


(a) Twitter

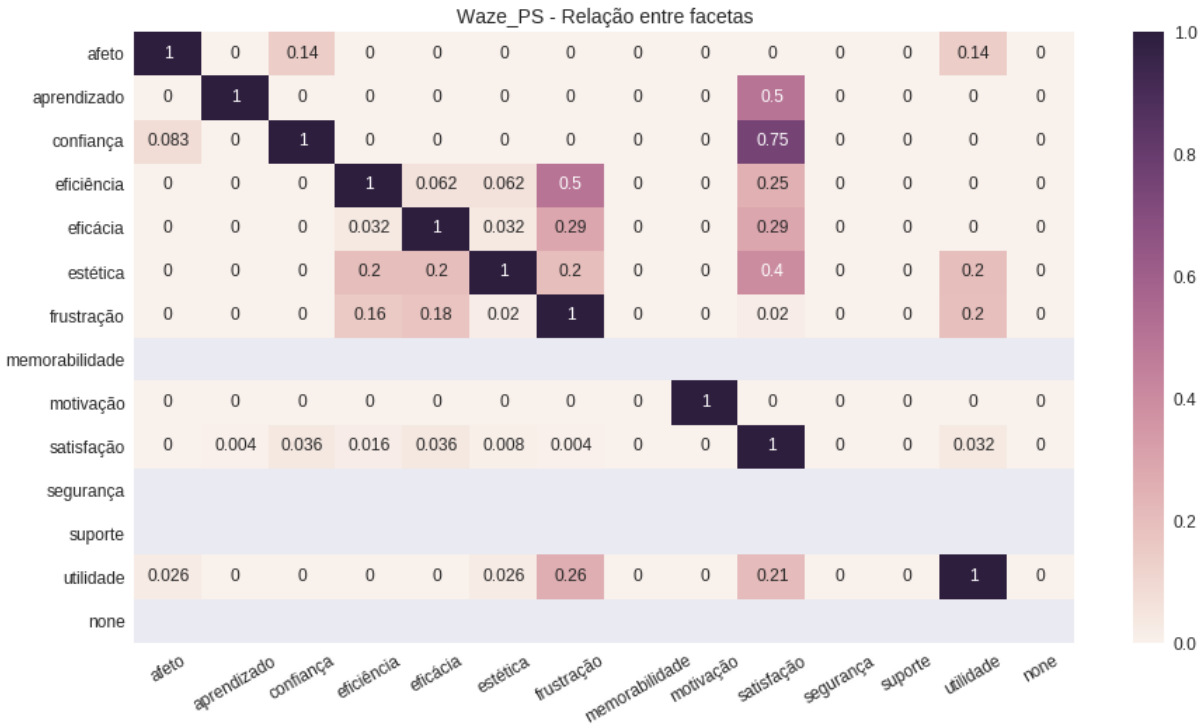


(b) PlayStore – Google Maps

Figura 14 – (continuação) Relação entre facetas para cada *dataset*.

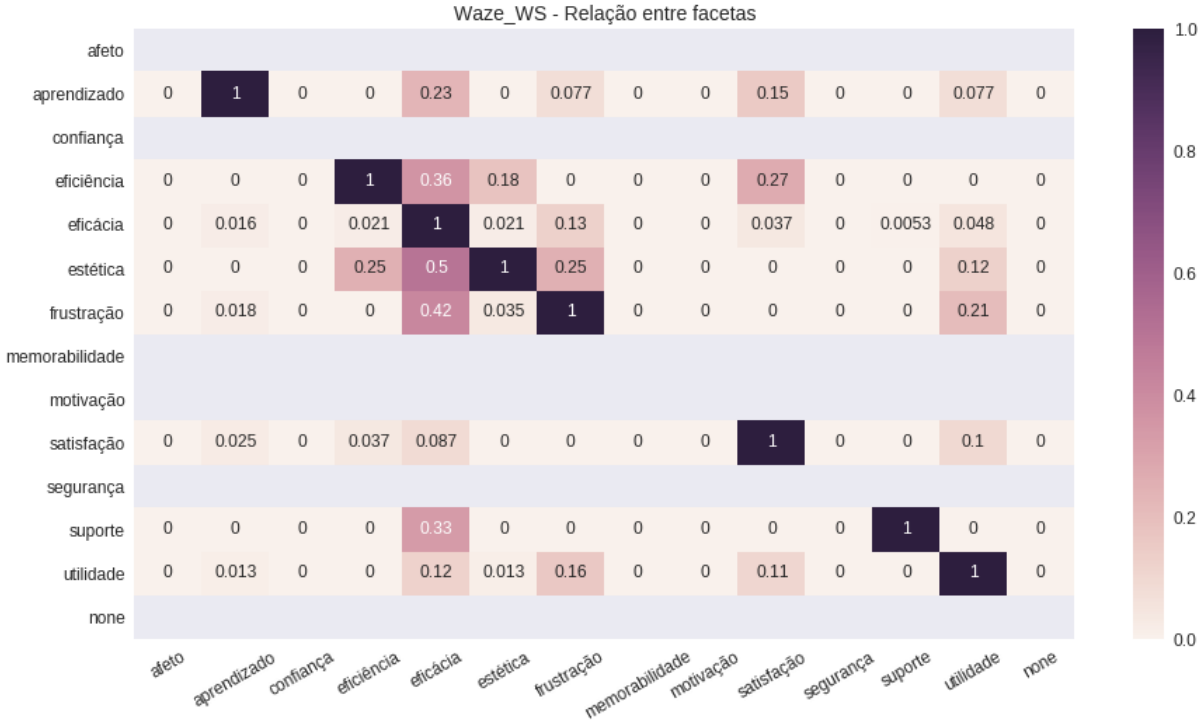


(a) ReclameAqui – Spotify

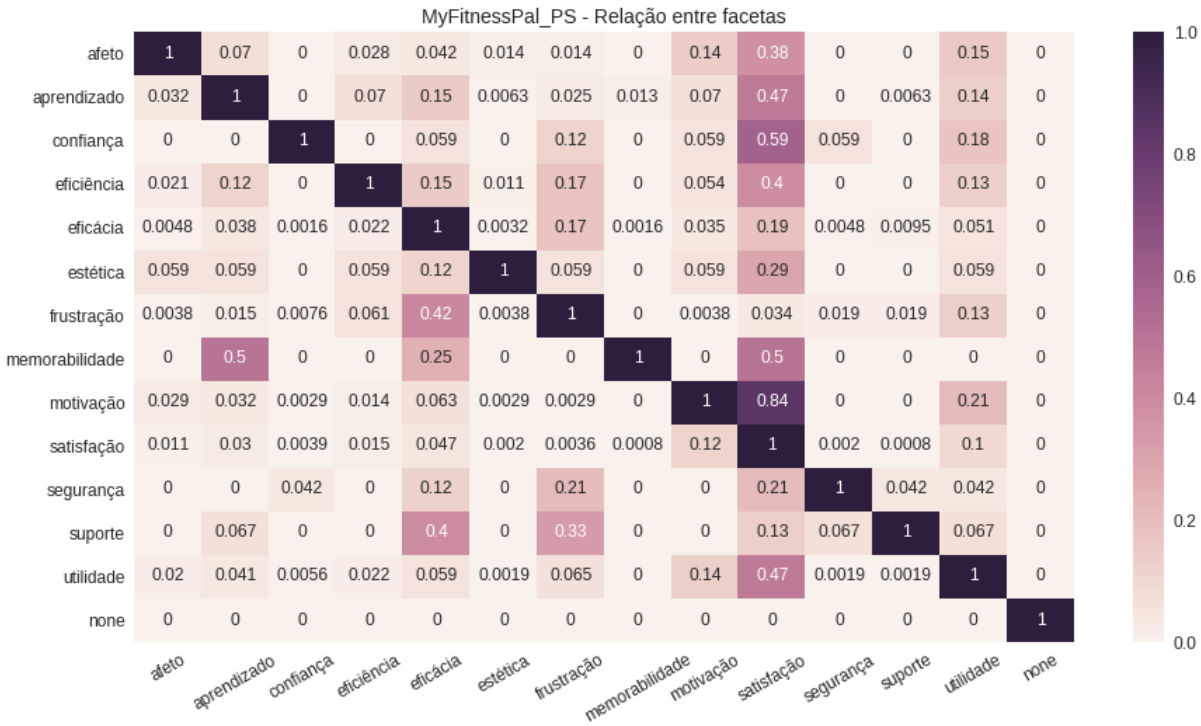


(b) PlayStore – Waze

Figura 14 – (continuação) Relação entre facetas para cada *dataset*.

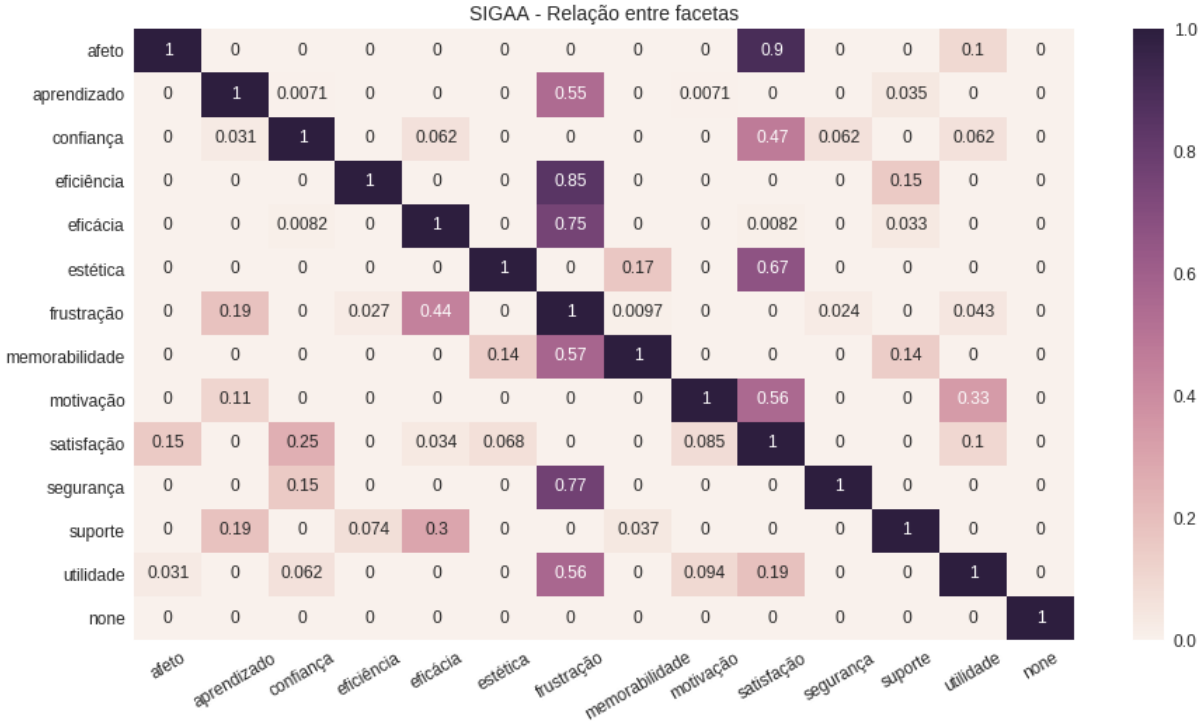


(a) WindowsStore – Waze



(b) PlayStore – MyFitnessPal

Figura 14 – (continuação) Relação entre facetas para cada *dataset*.



(a) SIGAA

Fonte: O autor (2018).

APÊNDICE B – RESULTADOS DA CLASSIFICAÇÃO COM NÚMERO REDUZIDO DE FACETAS.

Tabela 3 – Pontuações de cada algoritmo de aprendizado, destacando as maiores gerais e as maiores de Deep Learning, excluindo as facetas com quantidade menor que 1% da quantidade total.

Algoritmo	**Acurácia**	**Precisão**	**Recall**	**F1**	**Média**
Binary Relevance − SVM	**0,93909**	0,58854	**0,74422**	0,59526	0,7167775
Binary Relevance − Naive Bayes	0,93345	0,47201	0,72524	0,49065	0,6553375
Binary Relevance − SGD	0,93715	0,54692	0,73677	0,55701	0,6944625
Binary Relevance − Random Forest	0,93312	0,58879	0,7138	0,58637	0,70552
Classifier Chain − SVM	0,92972	**0,65579**	0,69241	**0,65896**	**0,73422**
Classifier Chain − Naive Bayes	0,93433	0,49786	0,72908	0,51755	0,669705
Classifier Chain − SGD	0,9279	0,63181	0,6787	0,64146	0,7199675
Classifier Chain − Random Forest	0,93028	0,61201	0,69887	0,60911	0,7125675
Label Powerset − SVM	0,93728	0,53534	0,74229	0,55091	0,691455
Label Powerset − Naive Bayes	0,93178	0,45759	0,71847	0,47848	0,64658
Label Powerset − SGD	0,93467	0,48458	0,73105	0,50096	0,662815
Label Powerset − Random Forest	0,93121	0,56084	0,70368	0,56898	0,6911775
Adapted Random Forest	0,93251	0,57328	0,70965	0,58137	0,6992025
MLkNN	0,93204	0,56118	0,71205	0,56904	0,6935775
NB−SVM	0,93771	0,59682	0,73496	0,59766	0,7167875
LSTM	0,93373	0,58876	0,71471	0,59315	0,7075875
LSTM−CNN	0,93309	**0,60379**	0,71059	**0,60558**	0,7132625
GRU	**0,93574**	0,59846	**0,72402**	0,59886	**0,71427**
GRU−CNN	0,93463	0,58146	0,71944	0,58175	0,70432
CNN	0,93548	0,57587	0,72396	0,57838	0,7034225

Fonte: O autor (2018).

Tabela 4 – Pontuações de cada algoritmo de aprendizado, destacando as maiores gerais e as maiores de Deep Learning, excluindo as facetas com quantidade menor que 5% da quantidade total.

Algoritmo	**Acurácia**	**Precisão**	**Recall**	**F1**	**Média**
Binary Relevance − SVM	0,89244	0,61694	0,75188	0,61985	0,7202775
Binary Relevance − Naive Bayes	0,88325	0,50591	0,734	0,52008	0,66081
Binary Relevance − SGD	0,88946	0,58675	0,74512	0,5917	0,7032575
Binary Relevance − Random Forest	0,88067	0,62449	0,72171	0,61548	0,7105875
Classifier Chain − SVM	0,88003	**0,68652**	0,71471	**0,68922**	**0,74262**
Classifier Chain − Naive Bayes	0,88394	0,54877	0,73539	0,56417	0,6830675
Classifier Chain − SGD	0,87512	0,66484	0,70196	0,67318	0,728775
Classifier Chain − Random Forest	0,86502	0,68288	0,68211	0,66615	0,72404
Label Powerset − SVM	**0,8934**	0,61844	**0,75665**	0,62729	0,723945
Label Powerset − Naive Bayes	0,87993	0,48543	0,72885	0,50214	0,6490875
Label Powerset − SGD	0,88878	0,55331	0,74782	0,56513	0,68876
Label Powerset − Random Forest	0,87675	0,59782	0,70886	0,60089	0,69608
Adapted Random Forest	0,87817	0,60472	0,71268	0,60908	0,7011625
MLkNN	0,88076	0,59169	0,72296	0,59653	0,697985
NB−SVM	0,88956	0,6225	0,74296	0,61918	0,71855
LSTM	0,88513	**0,65096**	0,72921	0,64713	0,7281075
LSTM−CNN	0,88428	0,64948	0,72763	**0,64926**	0,7276625
GRU	**0,88942**	0,65009	**0,73879**	0,64376	**0,730515**
GRU−CNN	0,88634	0,62815	0,73305	0,62344	0,717745
CNN	0,88706	0,62551	0,73491	0,62036	0,71696

Fonte: O autor (2018).

Tabela 5 – Pontuações de cada algoritmo de aprendizado, destacando as maiores gerais e as maiores de Deep Learning, excluindo as facetas com quantidade menor que 10% da quantidade total.

Algoritmo	**Acurácia**	**Precisão**	**Recall**	**F1**	**Média**
Binary Relevance − SVM	0,88245	0,65308	0,76437	0,65427	0,7385425
Binary Relevance − Naive Bayes	0,87244	0,55327	0,74516	0,56255	0,683355
Binary Relevance − SGD	0,87879	0,6187	0,75725	0,62154	0,71907
Binary Relevance − Random Forest	0,86763	0,65775	0,7323	0,64422	0,725475
Classifier Chain − SVM	0,87229	**0,71537**	0,74012	**0,71769**	**0,7613675**
Classifier Chain − Naive Bayes	0,87279	0,60563	0,74659	0,61729	0,710575
Classifier Chain − SGD	0,86806	0,70026	0,72983	0,70648	0,7511575
Classifier Chain − Random Forest	0,8531	0,70565	0,70244	0,6892	0,7375975
Label Powerset − SVM	**0,88344**	0,6558	**0,76841**	0,66233	0,742495
Label Powerset − Naive Bayes	0,86859	0,54144	0,73877	0,55341	0,6755525
Label Powerset − SGD	0,87942	0,60146	0,76064	0,60989	0,7128525
Label Powerset − Random Forest	0,86399	0,63428	0,72127	0,63607	0,7139025
Adapted Random Forest	0,86668	0,64351	0,72771	0,64485	0,7206875
MLkNN	0,8682	0,63333	0,73403	0,63581	0,7178425
NB−SVM	0,87847	0,64816	0,75481	0,6437	0,731285
LSTM	0,87449	0,67605	0,74526	0,67442	0,742555
LSTM−CNN	0,8742	0,68131	0,74519	**0,68054**	0,74531
GRU	**0,87899**	**0,68597**	**0,75412**	0,67895	**0,7495075**
GRU−CNN	0,87656	0,6692	0,74817	0,66051	0,73861
CNN	0,87589	0,6579	0,74738	0,65221	0,733345

Fonte: O autor (2018).

Tabela 6 – Pontuações de cada algoritmo de aprendizado, destacando as maiores gerais e as maiores de Deep Learning, excluindo as facetas com quantidade menor que 20% da quantidade total.

Algoritmo	**Acurácia**	**Precisão**	**Recall**	**F1**	**Média**
Binary Relevance − SVM	**0,8634**	0,66134	0,7627	0,66235	0,7374475
Binary Relevance − Naive Bayes	0,85611	0,5862	0,74966	0,5917	0,6959175
Binary Relevance − SGD	0,86093	0,65507	0,75812	0,65573	0,7324625
Binary Relevance − Random Forest	0,84578	0,67316	0,733	0,65733	0,7273175
Classifier Chain − SVM	0,85393	0,72496	0,74444	**0,72859**	**0,76298**
Classifier Chain − Naive Bayes	0,85546	0,64886	0,74873	0,65576	0,7272025
Classifier Chain − SGD	0,84888	0,71189	0,73492	0,71859	0,75357
Classifier Chain − Random Forest	0,83588	**0,72569**	0,71382	0,7099	0,7463225
Label Powerset − SVM	0,86315	0,67377	0,76206	0,67873	0,7444275
Label Powerset − Naive Bayes	0,85149	0,58706	0,74336	0,59359	0,693875
Label Powerset − SGD	0,86141	0,64151	0,76018	0,64723	0,7275825
Label Powerset − Random Forest	0,84315	0,66197	0,72367	0,66249	0,72282
Adapted Random Forest	0,8454	0,66572	0,72738	0,66654	0,72626
MLkNN	0,84651	0,64951	0,7322	0,64998	0,71955
NB−SVM	0,85901	0,64915	0,75371	0,64505	0,72673
LSTM	0,85527	0,69287	0,7456	0,69194	0,74642
LSTM−CNN	0,85427	0,69332	0,74424	**0,69464**	0,7466175
GRU	**0,85925**	**0,69391**	**0,75268**	0,68786	**0,748425**
GRU−CNN	0,8546	0,67824	0,74491	0,67073	0,73712
CNN	0,85601	0,67288	0,74816	0,66775	0,7362

Fonte: O autor (2018).