UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE CIÊNCIA DA COMPUTAÇÃO

ISAAC RAHEL MARTIM OLIVEIRA

ESTRATÉGIA PARA O PROBLEMA DE CLASSIFICAÇÃO DE POSTAGENS RELACIONADAS AO USO COM APRENDIZAGEM BASEADA EM REGRAS.

RUSSAS

Novembro, 2018

ISAAC RAHEL MARTIM OLIVEIRA

ESTRATÉGIA PARA O PROBLEMA DE CLASSIFICAÇÃO DE POSTAGENS RELACIONADAS AO USO COM APRENDIZAGEM BASEADA EM REGRAS.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus 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

Coorientadora: Profa. Dra. Marília Soares Mendes

RUSSAS

Novembro, 2018

ISAAC RAHEL MARTIM OLIVEIRA

ESTRATÉGIA PARA O PROBLEMA DE CLASSIFICAÇÃO DE POSTAGENS RELACIONADAS AO USO COM APRENDIZAGEM BASEADA EM REGRAS.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus 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

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Tiago Martins da Cunha

Universidade da Integração Internacional da Lusofonia Afro-Brasileira (UNILAB)

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)

O47e Oliveira, Isaac Rahel Martim.
Estratégia para o problema de classificação de postagens relacionadas ao uso com aprendizagem baseada em regras / Isaac Rahel Martim Oliveira. – 2018.
36 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. Aprendizagem baseada em regras. 2. IHC. 3. Aprendizagem de Máquina. I. Título.

CDD 005

AGRADECIMENTOS

Agradeço a todos que me acompanharam nessa longa jornada.

Aos amigos que moraram comigo neste tempo Afonso Matheus, Hugo Venâncio, Vinicius Almeida, Erik Almeida.

Aos que estão desde o primeiro semestre e tenho forte vínculo de amizade Carlos Victor, Marília Cristina, Mateus Oliveira, Igor Mendes, Marcos Alencar, Marcos Paulo.

A uma pessoa que não é da turma mas tenho grande orgulho de ter conhecido Elis Ionara.

Aos amigos do karatê que me ensinaram muito em especial o Renan Silva.

Aos professores que foram de grande apoio Alexandre Mattos, Daniel Siqueira, Filipe Maciel, Marília Mendes.

E principalmente a meu grande amigo de longa data Thomas Dillan Baltazar Mendonça por todos os anos de amizade juntos.

"What I am saying then is just because you don't know how you manage to be conscious, how you manage to grow and shape your body, doesn't mean that you're not doing it. Equally, if you don't know how the universe shines the stars, constellates the constellations, or galactifies the galaxies – you don't know but that doesn't mean that you aren't doing it just the same way as you are breathing without knowing how you breathe."

The essence of Alan Watts – Death part 2

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IHC	Interação Humano Computador
PRU	Postagem Relacionada ao Uso
NÃO-PRU	Postagem não Relacionada Ao Uso
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
MALTU	Modelo para Avaliação da interação em Sistemas Sociais a partir da Linguagem Textual do Usuário
UUX	Usabilidade e Experiência do Usuário
RI	Recuperação da Informação
SVM	*Support Vector Machine*
FCA	*Formal Concept Analysis*
TBL	*Transformation-Based Learning*
UUX-POSTS	Buscador de postagens relacionadas a UUX
ISO	*International Organization for Standardization*
PHP	*PHP: Hypertext Preprocessor*
SS	Sistemas Sociais
AJAX	*Asynchronous JavaScrip*t e *XML*
API	*Application Programming Interface*
TF-IDF	*Term Frequency–Inverse Document Frequency*
PLN	Processamento da Linguagem Natural
POS-Tagging	*Part-Of-Speech Tagging*
NER	*Named Entity Recognition*
NLTK	*Natural Language Toolkit*

SUMÁRIO

RESUMO

Com o grande e contínuo avanço tecnológico nos últimos anos se tornou possível a popularização de computadores e consequentemente os sistemas sociais somaram novos usuários, gerando uma grande quantidade de dados com a possibilidade de explorá-los. Existem diversas formas e técnicas específicas de Processamento da Linguagem Natural (PLN) e Aprendizado de Máquina para encontrar alguma informação útil proveniente destes dados. Este trabalho enfoca na área de Interação Humano Computador (IHC), especificamente na Usabilidade e eXperiência do Usuário (UUX) e Aprendizagem de Máquina. Em trabalhos estudados foi provado que os usuários de um sistema específico expressam como se sentem ao usar o mesmo, geralmente com o uso de postagens, a principal forma de interação da maioria dos sistemas sociais. Tais postagens são chamadas de Postagens Relacionadas ao Uso (PRUs). Outros trabalhos mostram uma metodologia que explora essas PRUs e uma ferramenta chamada UUX-POSTS, que presta suporte a extração e classificação da metodologia, porém com baixo índice de acerto. Este trabalho propõe o uso de uma estratégia baseada em regras para a classificação das postagens de uma rede social em PRU e NÃO-PRU e comparar os resultados com outras estratégias.

Palavras-chave: Aprendizagem Baseada em Regras. Classificação de Postagens. Aprendizagem de Máquina.

ABSTRACT

With the great and continuous technological advance in recent years it has become possible to popularize computers and consequently social systems have added new users, generating a large amount of data with the possibility of exploiting them. There are several specific forms and techniques of Natural Language Processing (PLN) and Machine Learning to find some useful information from these data. This work focuses on the Human Computer Interaction (IHC) area, specifically on Usability and User Experience (UUX) and Machine Learning. In the studies studied it has been proven that users of a specific system express how they feel when using the system, usually with the use of postings, the main interaction form of most social systems. Such postings are called Usage Related Posts (PRUs). Other works show a methodology that explores these PRUs and a tool called UUX-POSTS, which supports the extraction and classification of the methodology, but with a low success rate. This paper proposes the use of a rules-based strategy for classifying the posts of a social network in PRU and non-PRU and comparing the results with other strategies.

Key words: Rule-Based Learning. Posts Classification. Machine Learning.

1 INTRODUÇÃO

Com o grande e contínuo avanço tecnológico nos últimos anos se tornou possível a popularização de computadores e, de forma diretamente proporcional, os sistemas sociais aumentando a quantidade de dados produzidos e a possibilidade de explorá-los.

Diversos autores exploram esses dados no contexto de Interação Humano Computador (IHC). Durant e Smith (2006) analisam a polaridade de sentimentos em *logs Web*, Mendes, Furtado e Castro (2014) mostraram que os usuários expressam sua opinião sobre o sistema durante o uso dele por meio de Postagens Relacionadas ao Uso de sistemas (PRUs), postagens que descrevem a opinião do usuário sobre o sistema durante o uso, e Mendes (2015) apresenta uma ferramenta para extração e classificação das PRUs.

Com as informações provenientes das PRUs, os proprietários de sistemas podem ter uma noção do estado atual de seu sistema sob a visão do usuário, podendo identificar pontos fortes/fracos do sistema. Porém, como mostrado em Mendes e Furtado (2017) a ferramenta possui baixo acerto de classificação, o que pode acarretar em ruídos na visão do sistema.

A principal causa associada ao problema do baixo índice de acerto da ferramenta é a busca booleana, sistema antigo de RI e que está se mostrando pouco eficiente. No entanto, existem outros algorítimos de classificação como KNN e SVM e também algoritmos baseados em regras, que têm apresentado bons resultados como mostrado em (TARNAWSKI, FRACZEK, KRECICKI E JELEN, 2008). Por este motivo, esse trabalho visa investigar essa estratégia no contexto de avaliação de sistemas.

Esse trabalho de conclusão de curso propõe a implantação da classificação em PRU e NÃO-PRU com algoritmos de aprendizagem baseados em regras na classificação de postagens de usuários em redes sociais. É esperado aumentar a porcentagem de acerto na classificação binaria em PRU e NÃO-PRU.

2 TRABALHOS RELACIONADOS

Nessa seção são apresentados os trabalhos estudados e que reforçam a ideia deste trabalho, as seções são divididas em três áreas: opinião do usuário, análise de conteúdo e classificação textual. Há trabalhos que se encaixam em mais de uma área, estes estão na sua área principal.

2.1 Análise de conteúdo

O crescimento das redes sociais tornou a classificação manual dos dados gerados por elas, muito difícil. Por esse motivo Olowe, Gaber e Stahl (2013) apresentam uma gama de técnicas de Mineração de Dados e Aprendizagem de Máquina, usadas para diversos problemas como quantidade massiva de conteúdo, ruído na extração dos dados e relacionamento dos conteúdos.

Durant e Smith (2006) mostram uma classificação de sentimentos em mensagens de logs políticos da WEB de forma automática usando classificadores estatísticos Naive Bayes e Máquinas de Vetores de Suporte, Shein; Nyunt, (2010) oferecem uma combinação da classificação com SVM e uma ontologia baseada na Análise Formal de Conceito (FCA) com o objetivo de classificar a opinião dos usuários em relação ao sentimento em positivo, negativo ou neutro.

2.2 Classificação Textual

Do ponto de vista da aprendizagem baseada em regras, foram reunidos os trabalhos que representam as múltiplas aplicações da Aprendizagem Baseada em Transformação (TBL).

Ramshaw e Marcus (1995) demonstram que há muito tempo ele é usado para *Text Chunking,* Avinesh e Karthik (2007) mostram o uso em *POS-Tagging,* Fung et al., (2004) mostram o uso TBL com Entropia Máxima na execução de *Parsing* em bancos chineses onde há problemas muito negligenciados atuando em nível de caractere.

Como principal foco de estudo deste trabalho, Brill (1999) mostra uma combinação de algoritmos de aprendizado baseados em regras supervisionados e não supervisionados para desambiguação com alta precisão e compara com outros resultados de taggers.

2.3 Conclusão

Este trabalho se diferencia ao de Mendes (2015) principalmente no foco do trabalho, onde o deste trabalho é mostrar que a abordagem baseada em regras pode gerar bons resultados na classificação de postagens relacionadas ao uso e contribuir para a evolução da ferramenta UUX-POSTS, e se assemelha quanto ao contexto de análise em redes sociais e entrada dos usuários com linguagem natural.

Para esse intuito são usadas técnicas mostradas em (OLOWE; GABER; STAHL, 2013), alterando a amostragem da base de dados, e como mostradas na Seção 2.3 algumas das aplicações do TBL, esse trabalho usará principalmente POS-Tagging e aprendizagem de máquina.

3 OBJETIVOS

3.1 Objetivo geral

Este trabalho de conclusão de curso busca desenvolver uma estratégia para o problema de classificação de postagens relacionadas a usabilidade com aprendizagem baseada em regras de nível sintático.

3.2 Objetivos específicos

- Obter regras para a classificação de postagens em PRU e NÃO-PRU.
- Comparar com resultados que usam técnicas não baseada em regras em métricas como *Accuracy* e *Recall* com *f-measure*.

4 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo aborda e exemplifica quando possível. conceitos importantes para o entendimento do escopo deste trabalho. A Seção 4.1 define a Usabilidade e Experiência do Usuário (UUX) e suas principais métricas, seguida da Seção 4.2 que descreve o Processamento da Linguagem Natural, são definidos os Sistemas Sociais sob duas perspectivas de uso na Seção 4.3, a Seção 4.4 aborda a visão dos usuários através das Postagens Relacionadas ao Uso (PRU), com apoio da ferramenta UUX-POSTS, apresentada na Seção 4.5.

A partir da Seção 4.6 são abordados conceitos como a Recuperação da Informação e suas principais técnicas, que são apoiadas por estratégias e algoritmos de Mineração de Dados, apresentados na Seção 4.7 e finalmente na Seção 4.8 é apresentada a principal estratégia deste trabalho, o Aprendizado Baseado em Regras.

4.1 Usabilidade e Experiência do usuário.

Usabilidade é definida pela *ISO 9241-11:2018*[1] como uma medida em que um sistema, produto ou serviço pode ser usado antecipadamente ou não por usuários específicos para atingir objetivos específicos em um contexto específico de uso, o termo "específico" usado para usuários, objetivos e contexto de uso se refere a uma instância única dos mesmos, a usabilidade está sendo considerada, como Eficácia, Eficiência e Satisfação.

Nielsen, conceituado autor na área de Interação Humano Computador (IHC), define usabilidade em (NIELSEN, 1993) como atributo de qualidade que avalia conceitos sobre a utilização das interfaces do usuário.

Ressaltando, também, que a usabilidade não é um conceito único e unidimensional, mas sim uma combinação de múltiplos componentes, tradicionalmente: aprendizibilidade[2], eficiência, memorabilidade[3]e satisfação.

Experiência do usuário (UX) são as percepções e respostas que resultam do uso e/ou uso antecipado de um sistema, produto ou serviço, de acordo com a *ISO 9241-11:2018*[4]*;* Essas percepções e respostas tradicionalmente são denotadas por: emoções, crenças, preferências, confortos, comportamentos e realizações.

---

1 Consulta online disponível em: *https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en,* acesso em 04/05/2018.
2 O autor usa o termo "Learnability".
3 O autor usa o termo "Memorability".
4 Consulta online disponível em: *https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en,* acesso em 04/05/2018.

Há muitas definições para a UX e nenhuma é um consenso absoluto, de fato a experiência do usuário é um conceito controverso. Grande parte das definições se completam em termos de métricas para mensurar e quantificar a UX. Hartson e Pyla (2013) definem as seguintes métricas quantitativas para a experiência do usuário: performance inicial, performance de longo termo, capacidade de aprendizado[5], capacidade de retenção[6] e uso avançado de recursos.

Já Tullis e Albert (2013) dizem que todas as métricas para a experiência do usuário devem ser observáveis e agrega as seguintes: Eficácia, Eficiência e Satisfação. Sob outra perspectiva, Bevan (2009) retrata métricas mais focadas nas preferências específicas de cada usuário, tais: Conforto, Utilidade e Prazer.

Neste trabalho é adotada a definição da *ISO 9241-11:2018* tanto para usabilidade quanto para experiência do usuário.

4.2 Processamento da Linguagem Natural.

Jurafsky (2009) define como objetivo do Processamento da Linguagem Natural fazer com que os computadores realizem tarefas úteis envolvendo linguagem humana, tais como: possibilitar a comunicação da máquina com os humanos, melhorar a comunicação humano-humano ou simplesmente fazer um processamento útil de texto ou fala. Alguns exemplos mostram o quão abrangente a área é, abordando desde problemas antigos como contagem de palavras, hifenização automática, até aplicativos de última geração, como resposta automatizada a perguntas na Web e tradução em tempo real de idiomas falados.

Neste trabalho é adotada principalmente a categoria da Sintaxe que estuda as relações estruturais entre palavras, analisando a relação de vizinhança de termos específicos em uma determinada sentença.

4.3 Sistemas Sociais.

Mendes (2015) define Sistemas Sociais pelo termo "*Groupware",* que propriamente é traduzido como "software para grupos", muitos autores oferecem sua definição, Lynch et al (1990) distinguem *groupware* dos demais softwares que suportam multiusuários principalmente pelo conceito de proximidade dos usuários do sistema, enquanto *groupware* faz com que o usuário saiba que ele faz parte de um grupo, a maioria dos software

5 O autor usa o termo "Learnability".
6 O autor usa o termo "Retainability".

multiusuários tenta ocultar a presença dos demais usuários. O conceito é ampliado e complementado por Ellis et al (1991) que os definem como sistemas baseados em computador que suportam grupos de pessoas envolvidas em uma tarefa comum, que fornecem uma interface para um ambiente compartilhado e podem ser ou não coordenados por alguém.

Duas grandes perspectivas de sistemas sociais são:

1) **Conteúdo** – Sistemas cujo foco está na disponibilização e consumo de conteúdo, existem grandes exemplos como: YouTube, Spotify e Wikipédia.

2) **Interação** – Sistemas que focam no inter e/ou intra[7] relacionamento de seus usuários, como: Facebook, Twitter e Tinder.

A rede social utilizada, no contexto deste trabalho, será o Twitter. A motivação desa escolha se dá por que a ferramenta UUX-POST já usa a API do Twitter, devido também a facilidade de coleta dos dados.

4.4 Postagem relacionada ao uso (PRU).

É muito comum que os SS tenham a postagem como forma de interação entre seus usuários. Mendes (2015) determina PRU como postagem em linguagem natural na qual o autor da postagem se refere ao sistema social que ele está usando no momento.

Assim, dividem-se as postagens de um SS em Postagem Relacionada ao Uso do sistema (PRU) ou Postagem Não relacionada ao Uso do sistema (NÃO-PRU), o Twitter que é o contexto de onde serão retiradas as PRUs para esse trabalho através da ferramenta UUX-POSTS disponibilizada em Mendes (2015).

4.5 UUX-POSTS.

> É uma ferramenta de extração e classificação de postagens relacionadas a Usabilidade e Experiência do Usuário (UX). (UUX Posts – Analisador e classificador de postagens em de sites de redes sociais, disponível em <http://uuxposts.russas.ufc.br/>, acesso em 05/05/18).

A UUX-POSTS pode ser usada para coleta das postagens em redes sociais, ou também o usuário pode fazer upload de um banco de postagens já coletadas em formato .csv para classificação na ferramenta.

---

7 Inter: sistemas que encorajam o relacionamento de seus usuários no próprio sistema.
Intra: sistemas que incentivam o relacionamento fora do sistema, seja na vida real ou outros sistemas, servindo como facilitador desse relacionamento.

A ferramenta dispõe de uma coleção de padrões de busca relacionados a Usabilidade e Experiência do Usuário definidos em Mendes (2015), o usuário ainda tem a liberdade de alterar ou criar seus padrões.

A Figura 1 ilustra a interface da ferramenta UUX-POSTS.

Figura – 1 Interface da ferramenta UUX-POSTS.

Fonte: <http://uuxposts.russas.ufc.br/> acesso em 16/10/2018

A política d privacidade[8] que rege a ferramenta diz que:

1) A ferramenta só coleta as postagens de domínio público.

2) Os dados coletados sobre os autores das postagens são inteiramente disponibilizados pelos mesmos no site da rede social.[9]

3) A ferramenta não coleta as imagens.

4) A ferramenta não disponibiliza nomes dos autores, ou citados nas postagens.

Mendes e Furtado (2017) apresentam, em seu artigo sobre a ferramenta, alguns detalhes do funcionamento, para extração e classificação das postagens a ferramenta usa o *"modelo Booleano"* de Recuperação da Informação (RI), que usa palavras-chave para extração.

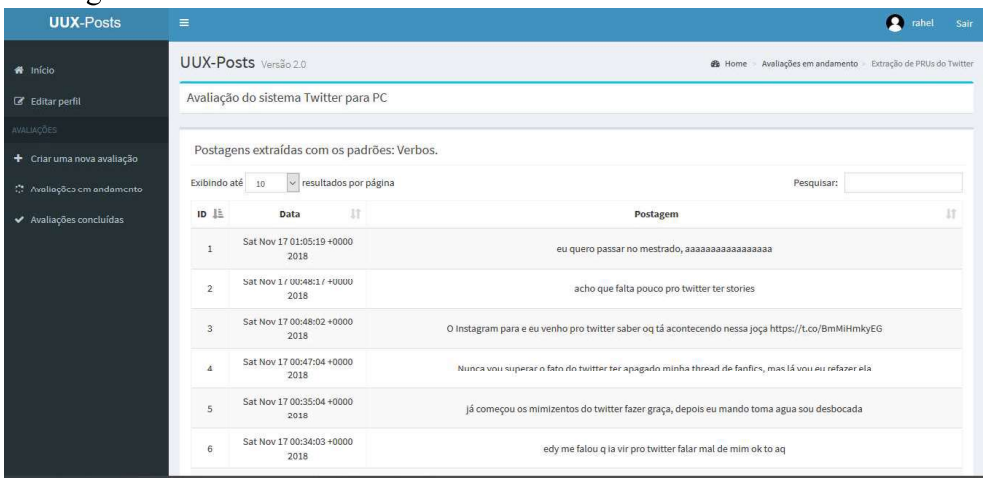
A UUX-POSTS está disponível para todos, tem código livre e em sua programação usa PHP, Javascript e AJAX (Asynchronous JavaScript e XML) para interface, classificação e chamadas as API's.

Essa será a ferramenta utilizada para coleta das postagens usadas como base dos algoritmos deste trabalho.

8 Dados retirados de <http://uuxposts.russas.ufc.br/> acesso em 05/05/18.
9 Dados como: idade, sexo e localização.

A Figura 2 apresenta os resultados de uma busca na ferramenta.

Figura – 2 Resultados de uma coleta do Twitter na UUX-POSTS.



Fonte: <http://uuxposts.russas.ufc.br/> acesso em 16/10/2018

Este trabalho visa também contribuir com a evolução da UUX-POSTS fornecendo outra forma de classificação para as postagens em PRU e NÃO-PRU com resultados melhores.

4.6 Recuperação da informação (RI).

Larson (2009) discute a Recuperação da Informação em termos gerais e amplos, e define a RI como os métodos para encontrar um material, geralmente documentos em forma de texto, e em computadores, de interesse com natureza não estruturada que satisfaz uma necessidade de informação em meio à grandes coleções.

Existem diversas técnicas de RI disponíveis para o estudo e uso, procedimentos mais complexas retornam resultados mais precisos, uma maior quantidade de resultados ou ainda com mais metadados,

Essa Seção se aprofunda em 3 técnicas específicas, as quais o modelo booleano, modelo vetorial e probabilistico:

1) **Modelo Booleano:**

Concebido da ideia da lógica booleana de George Boole usando os 3 operadores básicos: **AND**, **OR** e **NOT** para formular as operações de busca.

Chu (2003) ajuda a definir o modelo booleano, e seu procedimento. Os documentos da coleção são indexados por termos, um exemplo de indexação:

Tabela 1 – exemplo de documento não indexado.

"Eu odeio o Twitter, por n motivos."

Fonte: Elaborada pelo autor, 2018

Tabela 2 – documento anterior indexado por termos.

"Eu"	"odeio"	"o"	"Twitter"	"por"	"n"	"motivos"

Fonte: Elaborada pelo autor, 2018

A expressão de busca deve ser definida usando os operadores booleanos **AND**, **OR** e **NOT**, e os termos específicos como: *"Twitter **AND** erro"*. Ao aplicar a função de busca, são retornados os *t* documentos que atendem a expressão de busca.

2) **Modelo Vetorial:**

O modelo vetorial representa cada documento específico da coleção como vetores em um espaço vetorial comum, e essa representação é importante justamente para o desenvolvimento da lógica do modelo, e classificação, ranqueamento e agrupamento de documentos, afirmam Manning, Raghavan e Schütze (2009).

A indexação de cada documento para vetor é feita com a técnica Frequência de Termo e Frequência Inversa do Documento, TF-IDF (*Term Frequency – Inverse Document Frequency*) na qual um documento que possui mais menções a um termo específico de interesse do usuário tem mais importância para ele, logo deve ter uma pontuação mais alta (MANNING, RAGHAVAN e SCHÜTZE 2009). O peso final do documento é calculado pela soma do TF-IDF de todos os termos do documento podendo, então ser definido um ranking de documentos.

Após a indexação, é criado um vetor para cada documento da coleção, então é possivel visualizar a coleção de documentos como um espaço vetorial de N dimensões, onde N representa a quantidade de termos na coleção e, enfim comparar sua similaridade com outros vetores.

A similaridade de 2 vetores é calculada pelo cosseno do ângulo entres eles:

$$similaridade(d_x, d_y) = \cos(\theta_{d_x, d_y})$$

Forma-se um ranking dos documentos, onde aqueles que tem maior similaridade com a expressão de busca ficam em posições mais altas.

3) **Modelo Probabilístico:**

Para entender o modelo probabilístico um conceito importante para discussão é o de *Relevance Feedback*, Manning, Raghavan e Schütze (2009) relatam um problema muito

comum na RI, o de uma palavra poder ser interpretada de múltiplas formas diferentes, um exemplo clássico é o termo “manga” que pode significar uma fruta ou uma parte da roupa, outro exemplo, mais relacionado a este trabalho é o termo “erro”, onde há duas interpretações dependendo de seu uso:

1) “O Twitter está dando muito erro ultimamente”, é um exemplo de documento com ocorrência do termo “erro” do interesse do trabalho.

2) “O maior erro da minha vida foi entrar no Twitter”, é um exemplo de documento com ocorrência do termo “erro” que não há relevância para o trabalho.

Para resolver esse problema, é necessário que o avaliador do sistema nos diga qual uso é mais relevante para seu objetivo específico, e com a resposta do usuário é possivel consertar os pesos TF-IDF dos documentos para ranquear mais alto aqueles documentos de interesse do usuário.

Assim, pode-se construir uma base para “prever” (baseando-se em probabilidades condicionais), se um documento é relevante para o usuário ou não, denotando assim uma nova estratégia para RI que se destacam dois exemplos específicos o Naive Bayes e as Máquinas de vetores de suporte

4.7 Métricas

Algumas métricas para quantificação são conhecidas como clássicas e demonstram diferentes pontos de vista dos resultados obtidos, as principais exploradas neste trabalho foram **Acurácia, Cobertura e Precisão**.

Acurácia é a porcentagem de acerto bruta, definida como

$$Acurácia = \frac{ClassificaçõesCorretasPRU + ClassificaçõesCorretasNAOPRU}{TruePRU + TrueNAOPRU} * 100$$

Cobertura ou *Recall* pode ser conhecido por taxa de sensibilidade revela a porcentagem de falsos negativos, e a completude dos dados recuperados.

$$Cobertura\ PRU = \frac{ClassificaçõesCorretasPRU}{TruePRU} * 100$$

$$Cobertura\ NAO-PRU = \frac{ClassificaçõesCorretasNAOPRU}{TrueNAOPRU} * 100$$

Precisão representa a fração de postagens classificadas corretamente

$$Precisão\ PRU = \frac{ClassificaçõesCorretasPRU}{TodasClassificaçõesPRU} * 100$$

$$Precisão\ NAO-PRU = \frac{ClassificaçõesCorretasNAOPRU}{TodasClassificaçõesNAOPRU} * 100$$

4.8 Mineração de Dados.

A mineração de dados consiste em explorar uma grande coleção de dados procurando por padrões ou regras que possam ser convertidas em alguma informação útil para o proprietário da coleção.

Há clássicos exemplos de uso da mineração de dados para obter melhor desempenho em negócios, empresas provedoras de serviços podem usar mineração de dados armazenando os dados das transações de seus clientes e se verificado uma probabilidade do cliente abandonar o serviço, oferecer ofertas de interesse do cliente, como descontos ou promoções.

Como exemplo no contexto de uso deste trabalho na recuperação das postagens do Twitter a fim de inferir o estado atual da rede social a partir da visão e postagens dos usuários.

Hand, Smyth e Mannila (2001) oferecem sua definição de mineração de dados como:

> A mineração de dados é a análise de conjuntos de dados observacionais (geralmente grandes) para encontrar relacionamentos insuspeitos e para resumir os dados de maneiras novas que são compreensíveis e úteis para o proprietário dos dados. Os relacionamentos e resumos derivados de um exercício de mineração de dados são geralmente chamados de modelos ou padrões. Exemplos incluem equações lineares, regras, clusters, gráficos, estruturas de árvore e padrões recorrentes em séries temporais (Hand; Smyth; Mannila, 2001, pg 85).

A mineração de dados pode ser dividida em campos de acordo com o objetivo específico do usuário. Hand, Smyth e Mannila (2001) descrevem os objetivos mais comuns e sua parte da mineração de dados como: análise exploratória dos dados, modelagem descritiva, modelagem Preditiva (dividido em classificação e regressão), descoberta de padrões e regras e recuperação por conteúdo.

O contexto desse trabalho de recuperação da informação cabe nesse modelo na recuperação de textos por palavra-chave, na busca booleana.

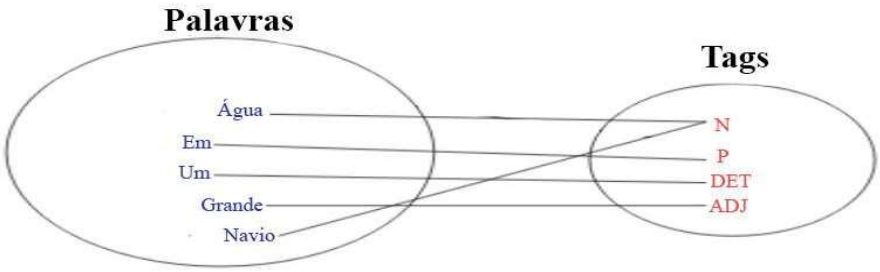
4.9 Aprendizado Baseado em Transformação (TBL)

Brill (1999) frente a desafios da análise automática da linguagem natural, formula pela primeira vez o TBL em seus estudos sobre *Part Of Speech Tagging (POS-Tagging)* e o problema da ambiguidade estrutural endêmica, significando que uma mesma sentença específica pode ter diversas análises possíveis, dependendo de sua gramática.

Essa abordagem guiada a erros é usada para aprender um conjunto ordenado de regras e é aplicado em múltiplos de estudo da linguagem natural:

1) ***POS tagging*** – cada palavra de uma sentença é analisada por sua função no contexto específico da sentença, ao contrário da habitual categorização lexical descontextualizada, e é definida uma tag, dependendo do conjunto de *tags* de POS adotado (VYAS et al, 2014). A Figura 3 apresenta um exemplo de *POS-Tagging.*

Figura – 3 Exemplo de POS-Tagging.



Fonte: Disponivel em <https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf>, POS tagging using UPennTreeBank II word tags.

2) ***Text Chunking*** – O *Text Chunking* envolve a divisão de sentenças em segmentos não sobrepostos com base em análises razoavelmente superficiais e como inclui a identificação das porções não recursivas de sintagmas nominais, ele também pode ser útil para outros fins, geralmente é usado como base para análises mais complexas (RAMSHAW e MARCUS, 1995) A Figura 4 ilustra um exemplo de *text chunking*.

Figura – 4 Exemplo de *Text Chunking*.

Sentence	He	reckons	the deficit	will narrow	to	1.8 billion
Chunk	⌣	⌣	⌣	⌣	⌣	⌣
Type	nominal	verbal	nominal	verbal	prep.	nominal
IOB2 Tag	B-NP	B-VP	B-NP I-NP	B-VP I-VP	B-PP	B-NP I-NP

Fonte: <https://www.maxwell.vrac.puc-rio.br/23812/23812_9.PDF> acesso em 10/05/18

3) ***Named Entity Recognition*** **(NER)** – O reconhecimento de entidades nomeadas tem como principal objetivo identificar e classificar entidades[10] em sentenças específicas definidas em língua natural. A figura 5 mostra um exemplo de NER.

10 As entidades podem ser pessoas, organizações, tempo, dinheiro, percentual etc.

Figura – 5 Exemplo de NER.

Fonte: CoNLL2003: http://www.cnts.ua.ac.be/conll2003/ner/

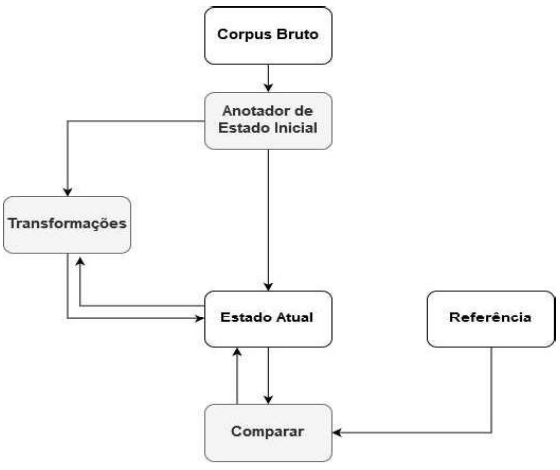
4) ***Prepositional Phrase Attachment*** – Merlo e Ferrer (2006) afirmam que a anexagem incorreta de frases preposicionais (PPs) é uma fonte de erros muito comuns no processamento da linguagem natural. O apego correto das PPs é necessário para construir uma árvore de análise que apoiará a interpretação adequada à análise dos constituintes da sentença, e apresenta o seguinte exemplo:

"Eu vi um homem com o telescópio."

No exemplo do autor há 2 interpretações possíveis dependendo da anexação da frase preposicional (PP) "com o telescópio", uma das análises seria que alguém viu um homem através do uso de um telescópio, e a outra seria que alguém viu um homem portando um telescópio. A figura 6 mostra os estados possíveis do TBL.

Brill (1999) descreve o esquema geral do TBL como:

Figura – 6 Exemplo dos estados do TBL.



Fonte: http://ccl.pku.edu.cn/doubtfire/TBL

I. O texto não anotado é passado por um anotador de estado inicial que pode ser de diferentes níveis de complexidade, dependendo do contexto de uso. O nível de complexidade varia desde uma abordagem simples como adotar tags NN e NP para todas as palavras, até complexos tagger's n-gram estocásticos.
II. A saída do anotador de estado inicial é comparada a um corpus de referência, e o sistema procura pela melhor transformação possível para o usuário, quantificada por uma função específica da aplicação do contexto de uso.
III. A transformação escolhida é aplicada ao estado atual em análise, gerando um novo estado ligeiramente alterado e é comparado ao corpus.
IV. O processo é iterado até não existir transformações que melhorem estado atual, a transformação que será aprendida é a última aplicada.
V. Ao final, tem-se uma sequência de transformações que podem ser aplicadas a qualquer texto que passe pelo anotador de estados inicial.

Brill (1999) mostra um exemplo do uso do TBL para desambiguação da palavra *race* para verbo ou substantivo, mostrando uma regra que a primeira palavra ou a segunda após um verbo modal (***will***) geralmente é um verbo, tendo como exceção quando a palavra também está logo a direita de um determinante (***the***) neste caso é um susbstantivo. A figura 7 mostra um exemplo de desambiguação da palavra race.

Figura – 7 Exemplo de desambiguação

(1) He **will race/VERB** the car.

(2) He **will** not **race/VERB** the car.

(3) When **will** the **race/NOUN** end?

Fonte: (BRIL ,1995).

5 EXPERIMENTO

Essa seção descreve o experimento realizado em razão dos objetivos deste trabalho. Na Seção 5.1 são apresentados os primeiros testes realizados, em seguida são mostrados os dados que serão utilizados e suas coletas na Seção 5.2. Em 5.3 é feita a limpeza de tais dados necessária para a técnica de POS-Tagging descrita em 5.4. A Seção 5.5 descreve a principal ideia deste trabalho e sua execução e comparação é feita em 5.6 e 5.7 respectivamente.

5.1 Testes iniciais

Para análise de desempenho e viabilidade foram realizados testes de *POS-Tagging* sobre bases padrões[11] como *Brown, Conll200* e *Treebank.* A linguagem usada foi Python 3.6.7 e a biblioteca para linguagem natural NLTK (*Natural Language Toolkit*).

Foram usadas 3 estratégias de POS-Tagging: N-*gram*[12]*, regexp e TBL*. Após aplicá-las nas 3 bases foi realizada uma validação e cálculo de métricas para comparação *Accuracy* e *Recall.*

5.2 Dados

Nessa etapa foram reunidos os dados que compõe a base que será usada para classificação neste trabalho, em trabalhos anteriores já havia sido coletado uma base de PRUs do Twitter com a ferramenta UUX-POSTS e validada manualmente por participantes do mesmo grupo de pesquisa, composta por 1938 PRUs.

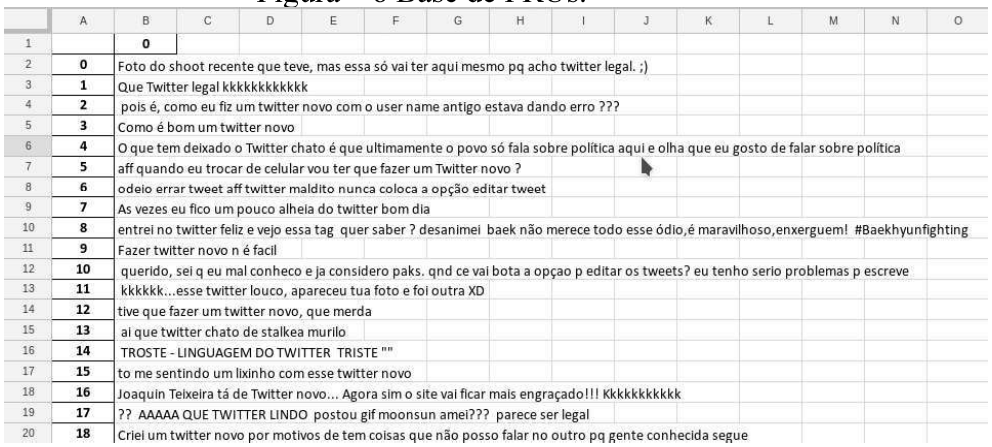
As Figuras 8 e 9 mostram parte das bases de PRU e NÃO-PRU coletadas e a Figura 10 mostra a ultima extração realizada para complementar as bases.

Foi realizada uma coleta de postagens do Twitter com a ferramenta UUX-POSTS, a extração se deu em 3 semanas, 1 extração a cada semana para evitar a repetição de postagens, após a extração os dados foram mesclados[13] e as postagens foram validadas manualmente resultando em 1360 NÃO-PRUs.

---

11 Padrões se refere a bases muito comuns, padronizadas e previamente classificadas para validação.
12 Foi testado com uni-gram, Bi-gram e Tri-gram.
13 Foi realizado retirada de cópias.

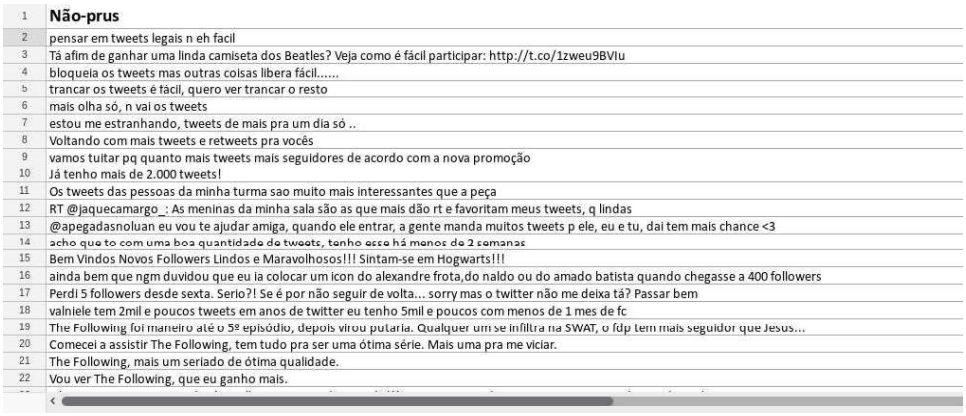
Figura – 8 Base de PRUs.



	A	B
1		0
2	0	Foto do shoot recente que teve, mas essa só vai ter aqui mesmo pq acho twitter legal. ;)
3	1	Que Twitter legal kkkkkkkkkkk
4	2	pois é, como eu fiz um twitter novo com o user name antigo estava dando erro ???
5	3	Como é bom um twitter novo
6	4	O que tem deixado o Twitter chato é que ultimamente o povo só fala sobre política aqui e olha que eu gosto de falar sobre política
7	5	aff quando eu trocar de celular vou ter que fazer um Twitter novo ?
8	6	odeio errar tweet aff twitter maldito nunca coloca a opção editar tweet
9	7	As vezes eu fico um pouco alheia do twitter bom dia
10	8	entrei no twitter feliz e vejo essa tag quer saber ? desanimei baek não merece todo esse ódio,é maravilhoso,enxerguem! #Baekhyunfighting
11	9	Fazer twitter novo n é facil
12	10	querido, sei q eu mal conheco e ja considero paks. qnd ce vai bota a opçao p editar os tweets? eu tenho serio problemas p escreve
13	11	kkkkkk...esse twitter louco, apareceu tua foto e foi outra XD
14	12	tive que fazer um twitter novo, que merda
15	13	ai que twitter chato de stalkea murilo
16	14	TROSTE - LINGUAGEM DO TWITTER TRISTE ""
17	15	to me sentindo um lixinho com esse twitter novo
18	16	Joaquin Teixeira tá de Twitter novo... Agora sim o site vai ficar mais engraçado!!! Kkkkkkkkkkk
19	17	?? AAAAA QUE TWITTER LINDO postou gif moonsun amei??? parece ser legal
20	18	Criei um twitter novo por motivos de tem coisas que não posso falar no outro pq gente conhecida segue

Fonte: Autor,2018

Figura – 9 Base de NÃO-PRUs.

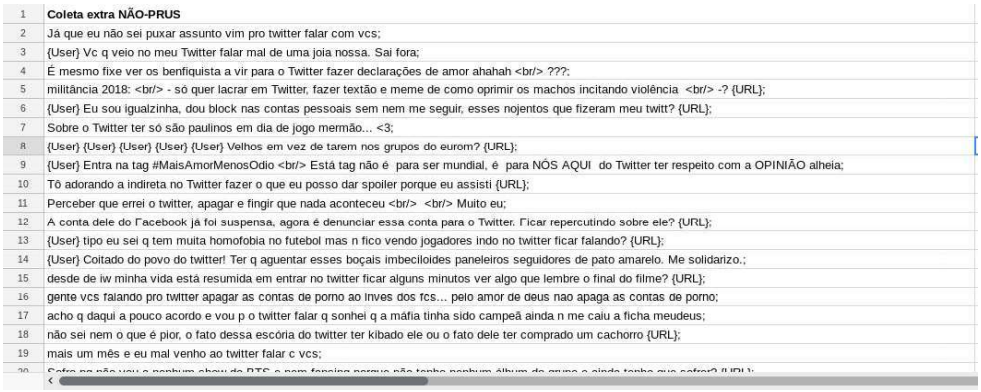


	Não-prus
2	pensar em tweets legais n eh facil
3	Tá afim de ganhar uma linda camiseta dos Beatles? Veja como é fácil participar: http://t.co/1zweu9BVIu
4	bloqueia os tweets mas outras coisas libera fácil.....
5	trancar os tweets é fácil, quero ver trancar o resto
6	mais olha só, n vai os tweets
7	estou me estranhando, tweets de mais pra um dia só ..
8	Voltando com mais tweets e retweets pra vocês
9	vamos tuitar pq quanto mais tweets mais seguidores de acordo com a nova promoção
10	Já tenho mais de 2.000 tweets!
11	Os tweets das pessoas da minha turma sao muito mais interessantes que a peça
12	RT @jaquecamargo_: As meninas da minha sala são as que mais dão rt e favoritam meus tweets, q lindas
13	@apegadasnoluan eu vou te ajudar amiga, quando ele entrar, a gente manda muitos tweets p ele, eu e tu, dai tem mais chance <3
14	acho que to com uma boa quantidade de tweets, tenho esse há menos de 2 semanas
15	Bem Vindos Novos Followers Lindos e Maravolhosos!!! Sintam-se em Hogwarts!!!
16	ainda bem que ngm duvidou que eu ia colocar um icon do alexandre frota,do naldo ou do amado batista quando chegasse a 400 followers
17	Perdi 5 followers desde sexta. Serio?! Se é por não seguir de volta... sorry mas o twitter não me deixa tá? Passar bem
18	valniele tem 2mil e poucos tweets em anos de twitter eu tenho 5mil e poucos com menos de 1 mes de fc
19	The Following foi maneiro até o 5º episódio, depois virou putaria. Qualquer um se infiltra na SWAT, o fdp tem mais seguidor que Jesus...
20	Comecei a assistir The Following, tem tudo pra ser uma ótima série. Mais uma pra me viciar.
21	The Following, mais um seriado de ótima qualidade.
22	Vou ver The Following, que eu ganho mais.

Fonte: Autor, 2018

Foi realizada mais uma extração de postagens visando aumentar a base de NÃO-PRUS, as postagens foram coletadas dado um tempo da extração anterior e validadas manualmente, resultando em 513 NÃO-PRUS.

Figura – 10 Base extra de NÃO-PRUs.



	Coleta extra NÃO-PRUS
2	Já que eu não sei puxar assunto vim pro twitter falar com vcs;
3	{User} Vc q veio no meu Twitter falar mal de uma joia nossa. Sai fora;
4	É mesmo fixe ver os benfiquista a vir para o Twitter fazer declarações de amor ahahah ???;
5	militância 2018: - só quer lacrar em Twitter, fazer textão e meme de como oprimir os machos incitando violência -? {URL};
6	{User} Eu sou igualzinha, dou block nas contas pessoais sem nem me seguir, esses nojentos que fizeram meu twitt? {URL};
7	Sobre o Twitter ter só são paulinos em dia de jogo mermão... <3;
8	{User} {User} {User} {User} {User} Velhos em vez de tarem nos grupos do eurom? {URL};
9	{User} Entra na tag #MaisAmorMenosOdio Está tag não é para ser mundial, é para NÓS AQUI do Twitter ter respeito com a OPINIÃO alheia;
10	Tô adorando a indireta no Twitter fazer o que eu posso dar spoiler porque eu assisti {URL};
11	Perceber que errei o twitter, apagar e fingir que nada aconteceu Muito eu;
12	A conta dele do Facebook já foi suspensa, agora é denunciar essa conta para o Twitter. Ficar repercutindo sobre ele? {URL};
13	{User} tipo eu sei q tem muita homofobia no futebol mas n fico vendo jogadores indo no twitter ficar falando? {URL};
14	{User} Coitado do povo do twitter! Ter q aguentar esses boçais imbeciloides paneleiros seguidores de pato amarelo. Me solidarizo.;
15	desde de iw minha vida está resumida em entrar no twitter ficar alguns minutos ver algo que lembre o final do filme? {URL};
16	gente vcs falando pro twitter apagar as contas de porno ao inves dos fcs... pelo amor de deus nao apaga as contas de porno;
17	acho q daqui a pouco acordo e vou p o twitter falar q sonhei q a máfia tinha sido campeã ainda n me caiu a ficha meudeus;
18	não sei nem o que é pior, o fato dessa escória do twitter ter kibado ele ou o fato dele ter comprado um cachorro {URL};
19	mais um mês e eu mal venho ao twitter falar c vcs;
20	Sofro na não vou a nenhum show do BTS e nem fansign porque não tenho nenhum álbum do grupo e ainda tenho que sofrer? {URL};

Fonte: Autor, 2018

Ao fim, tem-se a base final com 1938 PRUs e 1873 NÃO-PRUs.

5.3 Limpeza

Nessa fase foi realizada a execução de algoritmos de PLN visando retirar ruídos da base, melhorando a confiabilidade dos dados como retirada de *stopwords*, retirada de palavrões, redução ao radical e retirada de termos irrelevantes com TF-IDF. Foi também normalizada[14] a base para o treinamento dos métodos de classificação SVM e Naive.

Um exemplo da fase de limpeza dos dados retirado da base das postagens é:

1) “odeio errar tweet, aff twitter maldito nunca coloca a opção editar tweet !!!”

2) “odeio errar tweet twitter maldito nunca editar tweet”

Em 2 foram retirados os *stopwords* basicos e pontuações.

5.4 POS-Tagging

Neste passo foi efetuada a execução de POS-Tagging nas bases de PRU e NÃO-PRU, o algoritmo usado foi implementado em Python 3.6.7 usado a biblioteca NLTK e a estratégia TBL com o etiquetador BrillTagger e o tagset ‘universal’ do NTLK.

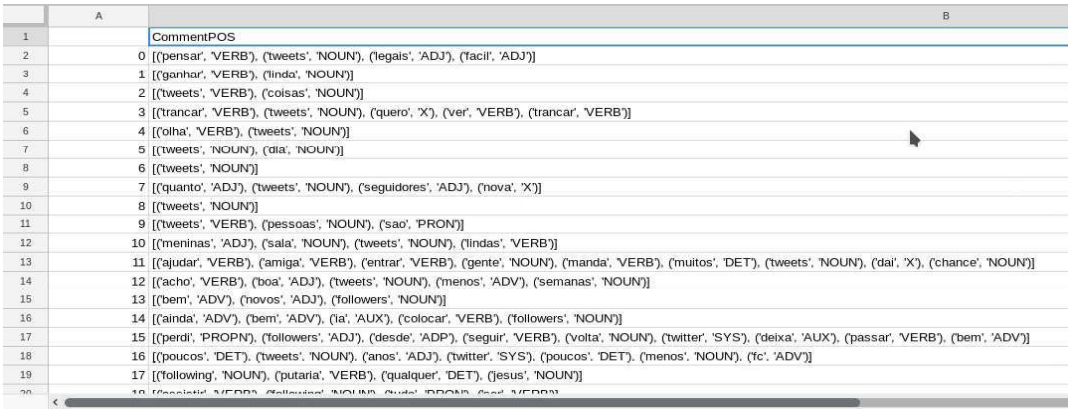
Figura – 11 Base de PRUs taggeada.

	A	B
1		CommentPOS
2	0	[('foto', 'VERB'), ('ter', 'AUX'), ('aqui', 'ADV'), ('acho', 'VERB'), ('twitter', 'SYS'), ('legal', 'NOUN')]
3	1	[('twitter', 'SYS'), ('legal', 'ADJ')]
4	2	[('fiz', 'VERB'), ('twitter', 'SYS'), ('novo', 'ADJ'), ('user', 'NOUN'), ('antigo', 'ADJ'), ('dando', 'VERB'), ('erro', 'NOUN')]
5	3	[('bom', 'ADJ'), ('twitter', 'SYS'), ('novo', 'ADJ')]
6	4	[('twitter', 'SYS'), ('chato', 'NOUN'), ('ultimamente', 'ADJ'), ('povo', 'NOUN'), ('fala', 'VERB'), ('sobre', 'ADP'), ('aqui', 'ADV'), ('olha', 'AUX'), ('gosto', 'NOUN'), ('falar', 'VERB'), ('sobre', 'ADP')]
7	5	[('odeio', 'NOUN'), ('trocar', 'VERB'), ('celular', 'VERB'), ('ter', 'AUX'), ('fazer', 'VERB'), ('twitter', 'SYS'), ('novo', 'NOUN')]
8	6	[('odeio', 'NOUN'), ('errar', 'VERB'), ('tweet', 'NOUN'), ('twitter', 'SYS'), ('maldito', 'NOUN'), ('nunca', 'ADV'), ('editar', 'VERB'), ('tweet', 'NOUN')]
9	7	[('vezes', 'NOUN'), ('fico', 'ADJ'), ('pouco', 'ADV'), ('alheia', 'VERB'), ('twitter', 'SYS'), ('bom', 'ADJ'), ('dia', 'NOUN')]
10	8	[('entrei', 'VERB'), ('twitter', 'SYS'), ('feliz', 'ADJ'), ('vejo', 'NOUN'), ('tag', 'ADJ'), ('quer', 'VERB'), ('saber', 'VERB'), ('todo', 'DET')]
11	9	[('fazer', 'VERB'), ('twitter', 'SYS'), ('novo', 'ADJ'), ('facil', 'NOUN')]
12	10	[('querido', 'VERB'), ('sei', 'VERB'), ('mal', 'ADV'), ('editar', 'VERB'), ('tweets', 'NOUN'), ('problemas', 'NOUN')]
13	11	[('twitter', 'SYS'), ('louco', 'VERB'), ('apareceu', 'VERB'), ('foto', 'ADV'), ('outra', 'ADV')]
14	12	[('fazer', 'VERB'), ('twitter', 'SYS'), ('novo', 'ADJ'), ('merda', 'NOUN')]
15	13	[('ai', 'VERB'), ('twitter', 'SYS'), ('chato', 'NOUN')]
16	14	[('twitter', 'SYS'), ('triste', 'NOUN')]
17	15	[('sentindo', 'VERB'), ('twitter', 'SYS'), ('novo', 'ADJ')]
18	16	[('twitter', 'SYS'), ('novo', 'ADJ'), ('agora', 'ADV'), ('sim', 'ADV'), ('site', 'NOUN'), ('ficar', 'VERB')]
19	17	[('twitter', 'SYS'), ('lindo', 'VERB'), ('gif', 'NOUN'), ('amei', 'X'), ('parece', 'AUX'), ('ser', 'VERB'), ('legal', 'ADJ')]

Fonte: Autor

14 Normalização se refere a estrutura dos dados e como eles se dispõem na planilha.

Figura – 12 Base de NÃO-PRUs taggeada.



	A	B
1		CommentPOS
2	0	[('pensar', 'VERB'), ('tweets', 'NOUN'), ('legais', 'ADJ'), ('facil', 'ADJ')]
3	1	[('ganhar', 'VERB'), ('linda', 'NOUN')]
4	2	[('tweets', 'VERB'), ('coisas', 'NOUN')]
5	3	[('trancar', 'VERB'), ('tweets', 'NOUN'), ('quero', 'X'), ('ver', 'VERB'), ('trancar', 'VERB')]
6	4	[('olha', 'VERB'), ('tweets', 'NOUN')]
7	5	[('tweets', 'NOUN'), ('dia', 'NOUN')]
8	6	[('tweets', 'NOUN')]
9	7	[('quanto', 'ADJ'), ('tweets', 'NOUN'), ('seguidores', 'ADJ'), ('nova', 'X')]
10	8	[('tweets', 'NOUN')]
11	9	[('tweets', 'VERB'), ('pessoas', 'NOUN'), ('sao', 'PRON')]
12	10	[('meninas', 'ADJ'), ('sala', 'NOUN'), ('tweets', 'NOUN'), ('lindas', 'VERB')]
13	11	[('ajudar', 'VERB'), ('amiga', 'VERB'), ('entrar', 'VERB'), ('gente', 'NOUN'), ('manda', 'VERB'), ('muitos', 'DET'), ('tweets', 'NOUN'), ('dai', 'X'), ('chance', 'NOUN')]
14	12	[('acho', 'VERB'), ('boa', 'ADJ'), ('tweets', 'NOUN'), ('menos', 'ADV'), ('semanas', 'NOUN')]
15	13	[('bem', 'ADV'), ('novos', 'ADJ'), ('followers', 'NOUN')]
16	14	[('ainda', 'ADV'), ('bem', 'ADV'), ('ia', 'AUX'), ('colocar', 'VERB'), ('followers', 'NOUN')]
17	15	[('perdi', 'PROPN'), ('followers', 'ADJ'), ('desde', 'ADP'), ('seguir', 'VERB'), ('volta', 'NOUN'), ('twitter', 'SYS'), ('deixa', 'AUX'), ('passar', 'VERB'), ('bem', 'ADV')]
18	16	[('poucos', 'DET'), ('tweets', 'NOUN'), ('anos', 'ADJ'), ('twitter', 'SYS'), ('poucos', 'DET'), ('menos', 'NOUN'), ('fc', 'ADV')]
19	17	[('following', 'NOUN'), ('putaria', 'VERB'), ('qualquer', 'DET'), ('jesus', 'NOUN')]

Fonte: Autor

5.5 Estratégia

Nesse ponto foi realizada a estratégia de aprendizado baseado em regras desenvolvida para o trabalho, primeira na base de PRUs e NÃO-PRUs cada ocorrência da palavra “twitter” foi taggeada com PRU ou NAO_PRU de acordo se a postagem é uma PRU ou não. Para que essa seja a base de validação.

Para a base de treinamento, cada ocorrência da palavra “twitter” foi taggeada como NAO_PRU, assim aplicando o algoritmo do TBL esperamos que ele “aprenda” quando uma menção da palavra “twitter” deve ser modificada de NÃO_PRU para PRU baseando-se na base validada.[15]

Para isso, apenas as sentenças que possuem pelo menos uma ocorrência da palavra ‘twitter’ serão levadas em conta, a intenção é avaliar se a partir da localidade das palavras ao termo ‘twitter’ é possível encontrar regras que indiquem que aquela postagem está se referindo ao uso do twitter ou apenas uma menção isolada a rede social.

Como foi mostrado a viabilidade em (Brill, 1995), é possível fazer a desambiguação de um termo a partir do uso das palavras vizinhas a ele até um certo grau de vizinhança, obviamente para frases muito prolixas essa técnica se torna menos confiável.

Nossa motivação ao escolher a estratégia baseada em regras, foi que a ferramenta apresentada em (MENDES, 2015) já faz uso da classificação baseada em regras por meio da

15 As outras palavras da sentença seguem com o taggeamento do NLTK.

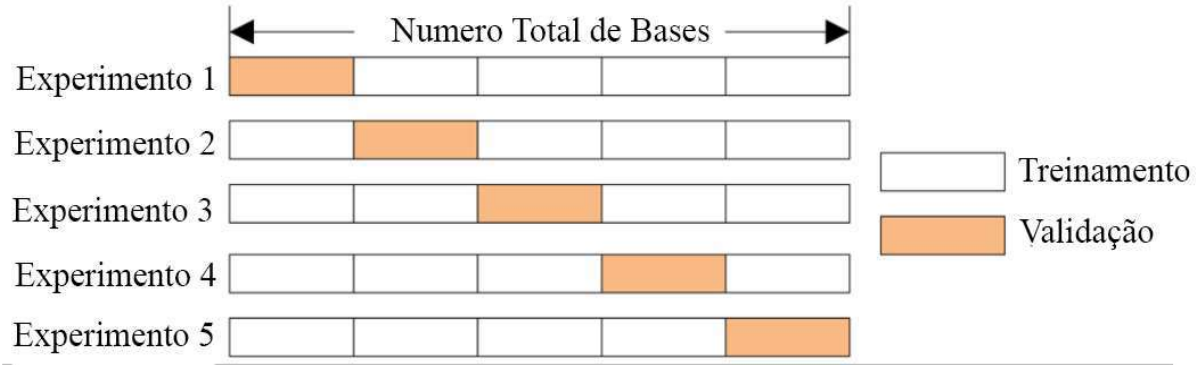
busca booleana, então a ideia é melhorar a forma com que a ferramenta classifica as postagens em PRU e NÃO-PRU, mantendo a estratégia baseada em regras porém gerando as regras automaticamente.

5.6 Classificação

Nesse ponto é realizada a classificação das postagens com 4 algoritmos: TBL com estratégia baseada em regras, Booleana, *Support Vector Machines* (SVM) e Naive Bayes, Como linguagem de programação usaremos Python 3.6.7 com a biblioteca NLTK.

E usando a validação cruzada com 5 *folds* com a proporção de 80/20, 80% da base para treinamento e 20% para validação da seguinte forma.

Figura – 13 Cross Validation



Fonte: Disponivel em *https://i.stack.imgur.com/1fXzJ.png*, acesso em 10/11/2018.

5.7 Comparação

Finalmente compararemos o resultado das classificações comparando em métricas de *Accuracy* e *Recall* com os algoritmos Naive Bayes Gaussiano, Regressão Logística, SVM e KNN, implementados em Python 3.6.7.

Todos os códigos estão disponíveis e mantidos em https://github.com/irahel/TCC-TBL-PRU-classification.

6 RESULTADOS

Essa seção apresenta os resultados obtidos, sua comparação com métodos clássicos de classificação como Naive Bayes Gaussiano, Regressão Logística, SVM e KNN, bem como a análise dos resultados e possíveis pontos a melhorar.

6.1 Resultados obtidos

A seguinte tabela mostra os resultados para a classificação das postagens em PRU e NÃO-PRU de acordo com as métricas mencionadas na seção anterior para a validação cruzada com 5 *folds*.

Foram analisadas um total de 2797 postagens sendo 2237 (80%) para treinamento e 560 (20%) para validação.

Tabela 4- Resultados TBL.

TBL	**PRU**	**NÃO-PRU**
Accuracy	83%	
Recall	91%	67%
Precision	85%	76%

Fonte: Elaborado pelo autor, 2018

As tabelas seguintes mostram os resultados da classificação com os métodos Naive Bayes Gaussiano, Regressão Logística, SVM respectivamente.

Tabela 5- Resultados Naive Bayes

Naive Bayes	**PRU**	**NÃO-PRU**
Accuracy	74%	
Recall	66%	91%
Precision	94%	56%

Fonte: Elaborado pelo autor, 2018

Tabela 6- Resultados Regressão Logística

Regressão Logística	**PRU**	**NÃO-PRU**
Accuracy	85%	
Recall	92%	71%
Precision	87%	81%

Fonte: Elaborado pelo autor, 2018

Tabela 7- Resultados SVM

SVM	PRU	NÃO-PRU
Accuracy	87%	
Recall	92%	76%
Precision	89%	82%

Fonte: Elaborado pelo autor, 2018

Tabela 8- Resultados KNN

KNN	PRU	NÃO-PRU
Accuracy	79%	
Recall	91%	52%
Precision	80%	75%

Fonte: Elaborado pelo autor, 2018

O algoritmo retorna uma lista de regras aprendidas, após o ranqueamento das regras e a retirada de regras que não influenciavam na escolha de "PRU" ou "NÃO-PRU" temos a seguinte lista:

As regras são ordenadas pelo *rank,* cada vez que uma regra é aprendida seu número de *rank* é incrementado e a regra se torna mais 'forte', o *rank* é usado para desempates de regras conflitantes, o maior *rank* define qual regra será aplicada.

Tabela 9- Lista de regras aprendidas

Rank	Regra	Condição
74	PRU->NAO_PRU	Sentença[1].tag = VERBO **and** Sentença[1] = "falar"
6	PRU->NAO_PRU	Sentença[1] = "ver"
27	NAO_PRU->PRU	Sentença[-1] = "pro"
24	PRU->NAO_PRU	Sentença[1].tag = VERBO **and** Sentença[1] = "fazer"
16	PRU->NAO_PRU	Sentença[1] = "porque"
13	NAO_PRU->PRU	Sentença[1] = "entrar"
12	NAO_PRU->PRU	Sentença[1].tag **or** Sentença[2].tag **or** Sentença[3].tag = ADJ
11	PRU->NAO_PRU	Sentença[-3] **or** Sentença[-2] **or** Sentença[-1] = "pro"
9	NAO_PRU->PRU	Sentença[-2].tag = PRU **and**

		Sentença[-1].tag = VERBO
7	PRU->NAO_PRU	Sentença[-1] = “vim”

Fonte: Elaborado pelo autor, 2018

A pontuação do ranqueamento é feito internamente pelo algoritmo, a regra define se a tag da palavra será trocada por exemplo “PRU→NAO_PRU” diz que a tag de PRU deve ser trocada para NAO_PRU se a condição for satisfeita.

A condição envolve diretamente a localidade das palavras para a criação da regra, por exemplo (Sentença[1] = “porque”) nos diz que se logo após a palavra (a localidade está representada entre colchetes) que está sendo analisada for a palavra “porque” a condição é verdadeira.

Por exemplo para essa regra a seguinte frase. “eu vim para o twitter porque gosto de falar de mim” teria inicialmente a tag de twitter como NÃO-PRU, manteria ela pois a condição da regra foi satisfeita, já que logo após a sentença ‘twitter’ vem a palavra ‘porque.

Como outro exemplo temos a regra (Sentença[-3] **or** Sentença[-2] **or** Sentença[-1] = “pro”) para a postagem “está dificil pro twitter implementar o botão de editar tweet logo?” terá a tag definida como PRU, pois antes da menção a ‘twitter” temos a palavra ‘pro’.

6.2 Conclusões

Podemos ver com o auxílio das tabelas apresentadas que o algoritmo usado com a abordagem de aprendizagem por regras obteve resultado melhores que Naive Bayes e KNN e tão bons quanto a Regressão Logística e SVM.

Então de acordo com os objetivos definidos temos que o objetivo geral

- Desenvolver uma estratégia para o problema de classificação de postagens relacionadas a usabilidade com aprendizagem baseada em regras de nível sintático.

Foi concluído visto a estratégia aplicada no trabalho, e os objetivos específicos

1. Obter regras para identificação de funcionalidades mencionadas em PRUs a partir da disposição das tags em uma PRU específica.
2. Comparar com resultados que usam técnicas não baseada em regras em métricas como *Accuracy* e *Recall* com *f-measure*

O objetivo 1 ficou fora do escopo do trabalho e o 2° objetivo foi concluído, adicionalmente foi possível gerar uma lista com as regras geradas automaticamente.

Então concluímos que é possível usar uma abordagem de aprendizagem baseada em regras para a classificação binaria em PRU e NÃO-PRU de postagens retiradas de uma rede social de forma satisfatória.

6.3 Possíveis Melhorias

Com a conclusão do trabalho temos algumas oportunidades de melhora que podem ser realizadas como por exemplo: O taggeamento foi feito de forma automática com o tagger da biblioteca NLTK em Python, logo há erros inevitáveis no taggeamento, uma forma de melhorar seria taggear manualmente com auxílio de um profissional com conhecimento em linguistica, ou testar a classificação alterando a tagset do tagger.

Outro ponto que pode resultar em melhora seria a aplicação de melhores tecnicas de mineração dos dados, alem do aumento da base.

Um ponto interessante a ser estudado é uma abordagem com lógica difusa (fuzzy) do algoritmo TBL.

6.4 Trabalhos Futuros

Esse trabalho abre oportunidade para muitos trabalhos futuros como a implantação do algoritmo de classificação na ferramenta UUX-POSTS de (mendes, 2015) visando melhorar o seu funcionamento.

Podemos também expandir a classificação para outras redes sociais visto que parte dela é feita a partir do taggeamento da palavra ‘twitter’ no nosso contexto, podemos, por exemplo, analisar posts do Facebook e usar a ocorrência de ‘facebook’ ou ‘face’.

Em próximos trabalhos podem ser estudado também o objetivo específico 1 de tentar identificar funcionalidades mencionadas a partir das regras.

REFERÊNCIAS

ADEDOYIN-OLOWE; MARIAM; GABER; MOHAMED; STAHL; FREDERIC. **A Survey of Data Mining Techniques for Social Media Analysis.** Journal of Data Mining and Digital Humanities. 2013.

BEVAN N. **What is the difference between the purpose of usability and user experience evaluation methods?**, UXEM'09 Workshop, INTERACT 2009, Uppsala, Sweden. 2009.

BRILL E.; POP M. **Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging. (eds) Natural Language Processing Using Very Large Corpora.** Text, Speech and Language Technology, vol 11. Springer, Dordrecht. 1999.

LYNCH; K.J.; SNYDER; J.M.; VOGEL; D.M; MCHENRY; W.K. **How Can We Make Groupware Practical?** Proceedings of the Conference on Human Factors in Computing Systems -CHI'90 (Apr. 1-5, Seattle, WA). ACM, N.Y., pp. 159-174. 1990.

DANIEL JURAFSKY AND JAMES H. MARTIN. **Speech and Language Processing (2nd Edition).** Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 2009.

DAVID J.; HAND; PADHRAIC SMYTH; HEIKKI MANNILA. **Principles of Data Mining.** MIT Press, Cambridge, MA, USA. 2001.

ELLIS C.A.; GIBBS, S.J.; REIN G.L. **Groupware: Some issues and experiences.** Communications of the ACM 34, 1. pp. 38-58. 1991.

ERIC BRIL. **Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging.** Comput. Linguist. 21, 4 December,pg 543-565. 1995.

HETING CHU. **Information Representation and Retrieval in the Digital Age.** Information Today, Inc., Medford, NJ, USA. 2003.

ISO 9241-11:2018 **Ergonomics of human-system interaction** -- Part 11: Usability: Definitions and concepts. 2018.

JAKOB NIELSEN. **Usability Engineering.** Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. Pg 23 – 36. 1993.

K. P. P. SHEIN; T. T. S. NYUNT. **Sentiment Classification Based on Ontology and SVM Classifier**, Second International Conference on Communication Software and Networks, Singapore, pp. 169-172. 2010.

KATHLEEN T. DURANT; MICHAEL D. SMITH. **Mining Sentiment Classification from Political Web Logs.** Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006.

MANNING, C.; RAGHAVAN; SCHÜTZE H. **Introduction to Information Retrieval.** Cambridge: Cambridge University Press. 2009.

MENDES Marília S. ; ELIZABETH S. FURTADO; MIGUEL F. DE CASTRO. **Do users write about the system in use?: an investigation from messages in natural language on Twitter.** Proceedings of the 7th Euro American Conference on Telematics and Information Systems (EATIS '14). ACM, New York, NY, USA, , Article 3 , 6 pages. 2014.

MENDES Marília S. . **MALTU – um modelo para avaliação da interação em sistemas sociais a partir da linguagem textual do usuário.** 199 f. Tese (Doutorado em Ciência da Computação)-Universidade Federal do Ceará, Fortaleza, 2015.

MENDES; FURTADO. **UUX-Posts: a tool for extracting and classifying postings related to the use of a system.** In Proceedings of the 8th Latin American Conference on Human-Computer Interaction (CLIHC '17). ACM, New York, NY, USA, Article 2, 8 pages. 2017.

MERLO Paola; E. E. FERRER. **The Notion of Argument in Prepositional Phrase Attachment.** Comput. Linguist. 32, 3 (September 2006), 341-378. 2006.

FUNG P; G. NGAI; Y. YANG; B. CHEN. **A maximum-entropy chinese parser augmented by transformation-based learning.** 3, 2 (June 2004), 159-168. 2004.

PVS, AVINESH; KARTHIK. **Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning**. 2007.

RAMSHAW, LANCE; MITCH Marcus. **Text Chunking Using Transformation-Based Learning. Proceedings of the Third Workshop on Very Large Corpora,**pp. 82–94. Association for Computational Linguistics, Somerset, New Jersey. 1995.

TULLIS Thomas TULLIS; ALBERT william. **Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics** 2ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 2013.

VYAS, YOGARSHI ET AL. “**POS Tagging of English-Hindi Code-Mixed Social Media Content.**” EMNLP. 2014.

WOJCIECH TARNAWSKI, MARCIN FRACZEK, TOMASZ KRECICKI, AND MICHAL JELEN. **Fuzzy rule-based classification system for computer-aided diagnosis in contact endoscopy imaging.** In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology (CSTST '08). ACM, New York, NY, USA, 457-463. 2008.