UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE SERRA TALHADA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Análise Comparativa de Métodos de Aprendizado Supervisionado para Mineração de Opinião dos Usuários da Plataforma de e-Participação VOTENAWEB

Por

Eliaquim Moreira Do Nascimento

Serra Talhada,
Junho/2019

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ELIAQUIM MOREIRA DO NASCIMENTO

Análise Comparativa de Métodos de Aprendizado Supervisionado para Mineração de Opinião dos Usuários da Plataforma de e-Participação VOTENAWEB

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Serra Talhada da Universidade Federal Rural de Pernambuco como requisito parcial à obtenção do grau de Bacharel.

Orientadora: Profa. Dra. Ellen Polliana Ramos Souza
Coorientador: Diego George da Silva Santos

Serra Talhada,
Junho/2019

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca da UAST, Serra Talhada - PE, Brasil.

N244a Nascimento, Eliaquim Moreira do

Análise Comparativa de métodos de aprendizado supervisionado para mineração de opinião dos usuários da plataforma de e-Participação VOTENAWEB / Eliaquim Moreira do Nascimento. – Serra Talhada, 2019.

47 f.: il.

Orientadora: Ellen Polliana Ramos Souza

Coorientador: Diego George da Silva Santos

Trabalho de Conclusão de Curso (Graduação em Bacharelado em Sistemas de Informação) – Universidade Federal Rural de Pernambuco. Unidade Acadêmica de Serra Talhada, 2019.

Inclui referências.

1. Sistemas de coleta automática de dados. 2. Processamento eletrônico de dados. 3. Pesquisa. I Souza, Ellen Polliana Ramos, orient. II. Santos, Diego George da Silva, coorient. III. Título.

CDD 004

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ELIAQUIM MOREIRA DO NASCIMENTO

Análise Comparativa de Métodos de Aprendizado Supervisionado para Mineração de Opinião dos Usuários da Plataforma de e-Participação VOTENAWEB

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendida e aprovada por unanimidade em XX/XX/XX pela banca examinadora.

Banca Examinadora:

Profa. Dra. Ellen Polliana Ramos Souza
Orientadora
Universidade Federal Rural de Pernambuco

Prof. Dr. Sérgio de Sá Leitão Paiva Júnior
Universidade Federal Rural de Pernambuco

Prof. Arthur Diego de Godoy Barbosa
Universidade Federal Rural de Pernambuco

Dedico este trabalho aos meus pais,

porque deram tudo de si para eu estar aqui.

AGRADECIMENTOS

Agradeço a todos que apoiaram esse tempo de dedicação, minha família em especial minha irmã Elma que ajudou-me sempre que pode, assim como meus professores todos que fortaleceram até antes tempo do ensino superior e em principal durante a realização deste trabalho. Agradeço com fervo a professora Ellen por ter tanta paciência.

"Dê a alguém asas, e ele pode querer voar muito perto do sol; dê a alguém o poder de ver o futuro, e ele pode ficar com medo de saber o que vai acontecer; e dê a alguém poderes ilimitados, e ele pode acreditar que nasceu para dominar o mundo.

(S. Kinberg, X-Men: Apocalypse)

"Tudo o que temos de decidir é o que fazer com o tempo que nos é dado."

(J. R. R. Tolkien, O Senhor dos Anéis)

RESUMO

Com a evolução das tecnologias da informação, passam a existir novos meios que visam promover uma sociedade mais democrática e participativa, como é o caso das plataformas de participação e colaboração eletrônicas, também conhecidas por *e-participation* e *e-collaboration*. Contudo, apesar de ser possível fornecer uma opinião na grande maioria dessas plataformas, tais opiniões não são analisadas e consideradas no processo de construção da proposta ou projeto de lei.É impossível para o ser humano compreender completamente todo o conteúdo em uma quantidade razoável de tempo, o que despertou um interesse na comunidade científica por sistemas capazes de extrair informações desse tipo de dado de forma automática. Mineração de opinião, também conhecida como análise de sentimento, é a área de estudo que analisa automaticamente sentimentos e opiniões das pessoas acerca de entidades, como produtos e serviços, expressos de forma não estruturada, como em texto, por exemplo. Neste sentido, este trabalho busca identificar o melhor conjunto de classificador versus técnica de de pre-processamento para análise das opiniões dos usuários de plataformas de e-participação e e-colaboração disponíveis para os cidadãos brasileiros. Como estudo de caso, para validação da aplicação, foram coletadas opiniões do portal VOTENAWEB, tendo em vista que o mesmo é bastante utilizado, além de permitir aos cidadãos postar comentários sobre um determinado projeto. Três algoritmos de aprendizado supervisionado, com diferentes técnicas de pré-processamento foram avaliadas que são tokenização, remoção de *stopwords*, N-grama, TF-IDF e a incorporação de palavras, a fim de obter a melhor configuração para mineração de opinião. O algoritimo de regressão linear obteve o melhor resultado com acurácia de 88,22% e f-medida de 87,07%, enquanto que o aprendizado profundo que o aprendizado profundo obteve acurácia de 84,96% e f-medida de 84,90%.

Palavras-chave: Mineração de Opinião. Análise de Sentimento. Língua Portuguesa. e-Participação. Aprendizado Profundo. VOTENAWEB.

ABSTRACT

With the help of information technologies, it is possible that the media and participatory media are recognized as democratic and participatory, as well as the participation and collaboration platforms, also known as e-participation and e-collaboration. However, although it is a problem, it is necessary to have a broader view on the perspectives, it is a process that is not analyzed and does not have a process of presentation or bill. It is impossible to be human completely in all content in a reasonable amount of time What has sparked an interest in the community of systems able to obtain additional information about the data type automatically. Opinion mining, also known as sentiment analysis, is an area of study that analyzes people's feelings and opinions about entities, such as products and services, unstructured formal expressions, such as text, for example. This study, this investigate the case of the class comparator and technique to pre-processing the reviews of users of topics-and-collaboration and e-collations available for the Brazilian citizens. As this article was published, it was published in the magazine Aplications of the portal VOTENAWEB, in view of what is widely used, in addition to allowing publications to comment on a set of projects. Supervised learning algorithms with data preprocessing techniques were evaluated: storpwords removal, keywords, N-gram, TF-IDF and word embedding, a set of instructions for keyword mining. . The linear regression algorithm was the best result with accuracy of 88.22% and f-measure of 87.07%, while the deep learning of the old meaning was accuracy of 84.96% and f-measure of 84.90%.

Keywords: Opinion Mining, Sentiment Analysis, Portuguese Language, e-Participation, Deep Learning, VOTENAWEB.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CGU	Conteúdo gerado pelo usuário
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
RL	Regressão Logística
SVM	*Support Vector Machines*
LSTM	*Long short-term memory*
RS	Remoção de *stopword*
RA	Remoção de acentuação
RN	Rede Neural
RP	Rede profunda

SUMÁRIO

1 Introdução

Neste capítulo, é apresentada a introdução deste trabalho. Na Seção 1.1, inicia uma contextualização do trabalho. A Seção 1.2 são expostas a motivação e a justificativa para o mesmo. Na Seção 1.3, demarcam-se os objetivos gerais e específicos. Por fim, a Seção 1.4 traz um resumo sobre a organização dos demais capítulos.

1.1 Contextualização

Nas últimas duas décadas, diversos autores têm realçado a importância de envolver o cidadão nas tomadas de decisões da administração pública, principalmente como elemento colaborativo na promoção de valores democráticos, tais como a capacidade de resposta e a co-responsabilização do cidadão perante uma nova sociedade (NELSON; WRIGHT et al., 1995; KING; FELTEY; SUSEL, 1998; FRANKLIN; EBDON, 2004; IRVIN; STANSBURY, 2004; FUNG, 2006).

Sendo assim, com auxílio das tecnologias da informação e comunicação (TIC), foram criados os conceitos de participação e colaboração eletrônicas, também conhecidos como e-participação e e-colaboração que, juntos, propõem um maior engajamento dos cidadãos no processo de tomada de decisão democrática (MACINTOSH, 2004).

O termo e-participação, derivado do inglês *e-participation*, é constituído pelo elemento 'e' de eletrônico, tendo uma associação com *e-government* e *e-governance*, e pelo termo 'participação', relacionando, assim, o uso de novas TIC, propondo uma maior participação cidadã na administração pública através dos meios eletrônicos (MACINTOSH, 2004).

A e-colaboração, por sua vez, surge com a premissa de 'construir em conjunto', utilizando plataformas nas quais as pessoas trabalham em projetos, de forma independente, para construção ou aprimoramento do todo, assim como acontece em projetos de desenvolvimento de software livre. Dentre as plataformas de e-colaboração, o portal e-Democracia da Câmara dos Deputados é um dos pioneiros neste conceito, tendo sido criado em 2009.

Já com relação às plataformas de e-participação, há várias iniciativas, destacando-se o portal e-Cidadania do Senado e o VOTENAWEB, que juntos, receberam mais de 5 milhões de

acessos entre os meses de março e maio de 2017 (SANTOS, 2017). Em ambas as plataformas, qualquer pessoa pode votar contra ou a favor de uma proposta e dar a sua opinião. Contudo, apesar de ser possível para o usuário opinar, tais opiniões não são analisadas nem consideradas no processo de construção da proposta ou do projeto de lei. Todas as análises são realizadas apenas pelo voto fornecido ao projeto como um todo, desconsiderando particularidades do texto.

Compreender o que as pessoas estão pensando ou suas opiniões é fundamental para a tomada de decisão, principalmente no contexto no qual essas pessoas exprimem seus comentários voluntariamente (ALVES et al., 2014). Entretanto, é impossível para o ser humano compreender um grande número de opiniões em uma quantidade razoável de tempo, o que despertou o interesse na comunidade científica por sistemas capazes de extrair automaticamente informações desse tipo de dado (BALAZS; VELÁSQUEZ, 2016).

Segundo Liu e Zhang (2012), mineração de opinião, também conhecida na literatura como análise de sentimento, é a área de estudo que analisa sentimentos, opiniões, avaliações, atitudes e emoções das pessoas acerca de entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos, expressos de forma não estruturada, como em um texto.

Nesse tipo de análise é realizada através da classificação de opinião de um documento, sentença ou característica em categorias, tais como: *positiva*, *negativa* ou *neutra*. De acordo com Pang e Lee (2008), esse tipo de classificação é referida na literatura como “polaridade de sentimento” ou “classificação de polaridade”.

Neste sentido, este trabalho tem como objetivo avaliar diferentes técnicas de mineração de opinião para analisar as opiniões dos usuários de plataformas de e-participação e e-colaboração disponíveis para os cidadãos brasileiros. Como pergunta norteadora da pesquisa levantamos a seguinte questão: que técnica de mineração de opinião mais se adéqua a esse contexto de estudo? Como estudo de caso, para validação da aplicação, foram coletadas opiniões do portal VOTE-NAWEB.A plataforma tem cerca 28.930 usuários ativos utilizado mensalmente tem e permitindo que os cidadãos postem seus comentários sobre um determinado projeto conforme Santos, Neto e Souza (2017).

1.2 Motivação e Justificativa

O aumento de plataformas que propõe uma melhor interação entre a gestão pública e os cidadãos, juntamente com as mídias sociais têm gerado o que é chamado de conteúdo gerado pelo usuário (CGU) na internet, fornecem informações que estão associadas sentimentos as opiniões, conforme Vitório et al. (2017).

Pra entender todo esse aglomerado de dados textuais que é gerado pelos usuários são necessárias ferramentas e técnicas específicas. Neste sentido, temos a mineração de opinião como sendo uma área multidisciplinar, cujo objetivo é entender as opiniões das pessoas de forma automática (PANG; LEE, 2008).

O entendimento aprimorado da análise sentimentos de acordo com (ZHANG; WANG; LIU, 2018), se faz necessário a buscar resultados melhores para ser possível impactar diferentes áreas assim como para negócios e gestão pública. As pessoas procuram opiniões diversas na tomada de decisão assim com as mais adequadas por parte dos gestores.

O aprendizado profundo por sua vez tem ganhado força para mineração de opinião pelos bons resultados obtidos por aglomerados de dados, não precisando de conhecimentos muito específicos da língua sendo assim aplicado técnicas especificas Collobert et al. (2011)

1.3 Objetivos

1.3.1 Objetivo Geral

O presente trabalho tem como objetivo identificar o melhor conjunto de classificador e técnicas de pre-processamento para análise das opiniões dos usuários de plataformas de e-participação e e-colaboração disponíveis para os cidadãos brasileiros.

1.3.2 Objetivos Específicos

Foram estabelecidos os seguintes objetivos específicos para atingir o objetivo geral:

- Expandir o corpus construído por Nascimento et al. (2017), contendo opiniões de cidadãos

brasileiros acerca de propostas ou projetos de lei;

- Desenvolver aplicação de mineração de opinião para tratamento dos dados, execução das classificações e avaliação, e;
- Identificar, executar e avaliar diferentes configurações de pré-processamento e algoritmos de aprendizagem;

1.4 Organização do trabalho

2 Referencial Teórico

Neste capítulo, é apresentada uma explanação acerca dos conteúdos utilizados neste trabalho. Na Seção 2.1, são apresentados conceitos de mineração de opinião. A Seção 2.3 explica o que é aprendizado de máquina, seus tipos e os algoritmos utilizados. Na Seção 2.3 apresenta abordagens de aprendizado utilizadas neste trabalho. Por fim, a Seção 2.5 apresenta os trabalhos relacionados.

2.1 Mineração de Opinião

Mineração de opinião, ou análise de sentimento, é considerada uma atividade de classificação de texto, já que discrimina a orientação de um sentimento, ou opinião, de um determinado texto em duas ou mais classes. Mineração de opinião tem sido realizada de várias formas no que se diz respeito à quantidade de classes, mas especialmente como binária ou ternária, *n*-ária na forma de estrelas, etc (RAVI; RAVI, 2015). Pesquisadores também têm levado em consideração vários tipos de emoção, como, por exemplo, as seis emoções universais: raiva, desgosto, medo, felicidade, tristeza e surpresa (PANG; LEE, 2008). O tipo de classificação usada neste estudo é uma classificação de rótulo único, onde cada opinião pertence a exatamente uma classe ou categoria (FELDMAN, 2013). As opiniões são classificadas em categorias como, por exemplo, 'positiva', 'negativa' ou 'neutra'. Esse tipo de classificação também é referida como 'polaridade de sentimento' ou 'classificação de polaridade'.

De acordo com um extenso levantamento realizado por (PANG; LEE, 2008), o termo 'mineração de opinião' (opinion mining) foi referenciado pela primeira vez por (DAVE; LAWRENCE; PENNOCK, 2003) num estudo realizado em 2003. Para esses autores, a ferramenta de mineração de opinião ideal deveria processar um conjunto de resultados de pesquisa para um determinado item, gerando uma lista de atributos e agregando opiniões acerca de cada um desses atributos. Muito da pesquisa subsequente, auto-identificada como mineração de opinião, se encaixa nessa definição, porém o termo também tem sido interpretado mais amplamente para incluir outros tipos de análise de texto (PANG; LEE, 2008).

Mineração de opinião costuma ser realizada em três diferentes níveis de análise: docu-

mento, sentença ou aspecto (FELDMAN, 2013). A classificação a nível de documento determina se a opinião de todo o documento é positiva, negativa ou neutra, por exemplo. A classificação a nível de sentença determina se uma determinada sentença expressa uma opinião positiva, negativa ou neutra. Já o nível de aspecto, ou característica, foca em todas as expressões de sentimentos presentes em um dado documento e o aspecto ao qual elas se referem. Neste estudo, foi levado em consideração a análise de sentimento a nível de documento.

As abordagens de mineração de opinião mais frequentemente utilizadas são as de aprendizado de máquina supervisionado e as baseadas em léxicos. Há também abordagens híbridas, que usam tanto aprendizado de máquina quanto léxicos (MEDHAT; HASSAN; KORASHY, 2014; PANG; LEE, 2008; RAVI; RAVI, 2015). Os métodos de aprendizado de máquina supervisionado aplicam algoritmos de classificação para aprender padrões subjacentes a partir de dados de exemplo com o objetivo de classificar novos dados não rotulados (BALAZS; VELÁSQUEZ, 2016). Para esses métodos, são necessários dois conjuntos de dados anotados: um para treinamento e outro para teste. Um bom número de algoritmos de aprendizado de máquina supervisionado têm sido utilizado para classificar opiniões. Dentre eles, se destacam o Naïve-Bayes e *Support Vector Machines* (SVM), que têm alcançado grande sucesso com mineração de opinião (RAVI; RAVI, 2015; SOUZA et al., 2016).

A abordagem baseada em léxicos, também conhecida como semântica ou simbólica, faz uso de palavras de opinião positivas, usadas para expressar estados desejados, e palavras de opinião negativas, usadas para expressar estados indesejados. Existem também frases de opinião e expressões idiomáticas que, juntas, são chamadas de léxico de opinião (MEDHAT; HASSAN; KORASHY, 2014). Três abordagens principais são usadas para construir léxicos de opinião: abordagem manual, que consume muito tempo; baseada em dicionários, onde um conjunto inicial, construído manualmente, é incrementado procurando os seus antônimos e sinônimos em corpora como a *WordNet* ou *thesaurus*; e baseado em corpus, que começa com uma lista semente de palavras de opinião para encontrar outras palavras de opinião em um grande corpus com orientações específicas de contexto.

Nos últimos anos conforme Zhang, Wang e Liu (2018), tem se destacado a utilização de aprendizado profundo para análise de sentimento, com essa crescente popularização tem sido um aprimorado devido a disponibilidade de poder de computacional pelos avanços nos hardwares, como exemplo o uso das GPUs,e a quantidade de bases para disponibilizadas para uso, assim tem demostrado o poder e a flexibilidade de aprender representações intermediárias. Assim, conseguindo números significantes de acertos.

2.2 Plataformas de e-Participação e e-Colaboração

O Brasil adota um regime político democrático, modelo de governo que vem sofrendo sinais de desgaste, como diz (FREITAS et al., 2015). Mas, com o surgimento e aprimoramento da Internet e a evolução das TICs, passam a existir novos meios que visam promover uma sociedade mais democrática e participativa, como é o caso da participação e colaboração eletrônicas ou *e-participation* e *e-collaboration.*

Depois do surgimento da web 2.0, termo cunhado em 2005 para se referir a uma nova geração de tecnologias que permite uma comunicação com maior nível de interatividade e colaboração através da Internet, oferecendo recursos e serviços para aplicações *online* e multimídia (BRESSAN, 2007), surgiram inúmeras plataformas como redes sociais, *blogs, chats, wikis* entre outras, incluindo as ferramentas de *e-participation* e *e-collaboration*.

Essas plataformas servem para auxiliar o cidadão a contribuir nas tomadas de decisão da administração pública através de *sites* e aplicativos, nos quais o cidadão pode opinar sobre um projeto de lei ou enviar suas ideias e sugestões de melhorias e leis, com objetivo de melhorar o meio em que vive.

Plataformas como Colab.re, e-Cidadania, Participa.br, dentre outras, disponibilizam mecanismos para que os cidadãos tomem conhecimento das ideias da administração pública, permitindo que os mesmos expressem sua aprovação ou desaprovação, trocando ideias com o poder público e com outros cidadãos.

A Tabela 2.1 apresenta um resumo das principais plataformas disponibilizadas para os cidadãos brasileiros. A coluna "Tipo", informa se a plataforma é de participação ou colaboração; a coluna "Status", informa a quantidade de acessos a suas páginas na Internet entre os meses de março a maio de 2017, e a Quantidade de downloads realizados na loja de aplicativos do Google até o mês de maio de 2017. Caso a plataforma não possua um aplicativo móvel é utilizado “n/a” para informar que essa categoria não se aplica àquela plataforma assim como apresentado no trabalho de (SANTOS; NETO; SOUZA, 2017).

Quadro 2.1 – Informações públicas das plataformas

Nome	Tipo	Status	Acesso	Downloads
e-Cidadania (Senado)	Participação	Ativo	5.060.000	-
e-Democracia (Câmara)	Colaboração	Ativo	38.710	1.000
Colab.re	Participação	Ativo	45.190	50.000
PortoAlegre.cc	Participação	Inativo	-	-
VOTENAWEB	Participação	Ativo	28.930	1.000
Mudamos	Participação	Inativo	180.920	100.000
Ouvidoria Cidadã da Câmara Municipal da cidade Vitória de Santo Antão	Participação	Ativo	5.000	500
Dialoga Brasil	Participação	Ativo	5.000	-
Participa.br	Colaboração	Ativo	36.080	1.000
Participa Maranhão	Participação	Ativo	15.530	-
Participa Cidadão (Bahia)	Participação	Ativo	0	50
Participa Campinas	Participação	Inativo	-	100
Avaaz	Participação	Ativo	6.710.000	-
Plamob Olinda	Participação	Ativo	0	-
SP 156	Participação	Ativo	54.860	10.000
Participa Cidadão (Cabo Verde-MG)	Participação	Ativo	29.700	100
Kit Urbano (18 Cidades brasileiras)	Participação	Ativo	-	10

Fonte: Elaborado por (SANTOS; NETO; SOUZA, 2017)

2.2.1 Portal VOTENAWEB

VOTENAWEB[1] é um portal de engajamento cívico apartidário que apresenta, de forma simples e resumida, os projetos de lei em tramitação no Congresso Nacional. Qualquer pessoa pode votar contra ou a favor das propostas e dar a sua opinião, conforme apresentado na Figura 2.1. O site fica encarregado de levar ao Congresso os resultados da participação popular. O objetivo do VOTENAWEB é aumentar a politização da sociedade, oferecer uma maneira fácil de acompanhar, votar e debater sobre o trabalho dos políticos, tudo isso num ambiente favorável ao diálogo entre parlamentares e cidadãos.

Além de contabilizar os votos dos cidadãos, como apresentado na Figura 2.1, ainda é possível que os internautas comparem seus votos entre si e com os dos deputados e senadores como mostrado na Figura 2.2. Por fim, a Figura 2.3 mostra um exemplo de comentário de um cidadão que, apesar de ter votado "Sim"para o projeto de lei como um todo, há partes em que o cidadão não concorda.

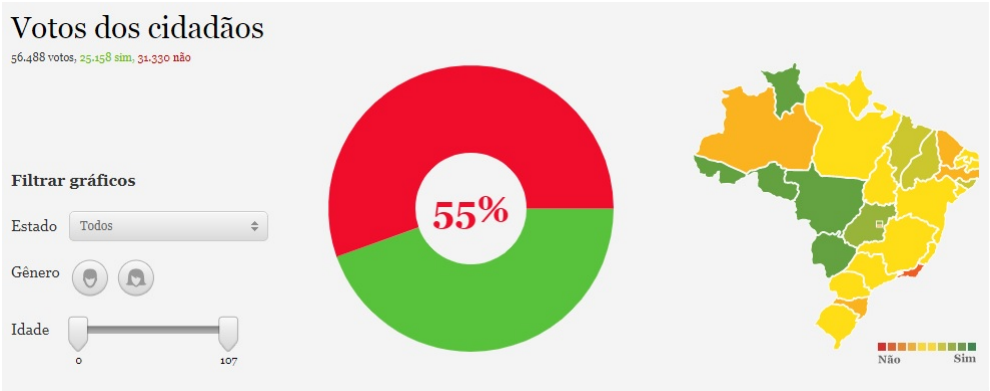
O portal "Política de Boteco"é a primeira iniciativa que leva aos bares do Brasil a discussão dos projetos de lei mais controversos do Congresso Nacional e entrega aos deputados

[1] http://www.votenaweb.com.br

Figura 2.1 – Resumo do projeto

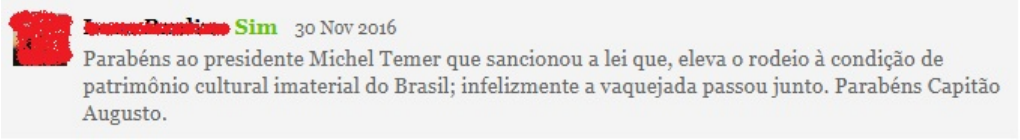
Fonte: Portal VOTENAWEB.

Figura 2.2 – Análise do voto dos cidadãos.



Fonte: Portal VOTENAWEB.

Figura 2.3 – Comentário do cidadão - Portal VOTENAWEB.



Fonte: Portal VOTENAWEB.

e senadores as discussões que acontecem no aplicativo. O projeto é fruto da parceria entre as empresas Webcitizen, através da plataforma VOTENAWEB, e Enox On-Life, com o objetivo de

atrair o interesse dos jovens que estão nos bares brasileiros para acompanhar os projetos de lei que estão em tramitação no Congresso, conhecer e discutir o conteúdo, compartilhar e opinar por meio de votos e comentários. Os projetos de lei do "Política de Boteco"e muitos outros projetos estão disponíveis no portal VOTENAWEB.

2.3 Aprendizado de Máquina

Como apresentado na Seção 2.1, técnicas de aprendizado de máquina são bastante utilizadas para mineração de opinião. Além disso, algoritmos de aprendizado de máquina têm se mostrado bastante úteis em diversos domínios, principalmente na mineração de dados (MITCHELL, 1997).

Aprendizado de máquina é a área que se preocupa em construir programas de computador que possam melhorar seus resultados a partir de experiências (MITCHELL, 1997). Logo, basicamente qualquer método que incorpore informação a partir de amostras de treinamento emprega aprendizagem e praticamente todos os problemas interessantes de reconhecimento de padrões são tão complicados que se deve considerar o uso de aprendizagem (DUDA; HART; STORK, 2000).

2.3.1 Tipos de Aprendizado de Máquina

Há três tipos de aprendizado de máquina: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço (DUDA; HART; STORK, 2000). Neste trabalho são utilizados algoritmos de classificação, baseados em aprendizado supervisionado.

No aprendizado supervisionado, o algoritmo, neste caso um classificador, recebe um conjunto de dados rotulados identificando a classe à qual eles pertencem (HAN; KAMBER, 2006). Neste trabalho, o classificador recebeu um conjunto de opiniões com suas respectivas polaridades e, a partir desses exemplos, pôde classificar novas opiniões. Para isso, são necessárias duas etapas: primeiramente, é realizada a etapa de treinamento, na qual o classificador é alimentado (treinado) com os dados rotulados para que ele possa "aprender"a partir deles; após o treinamento, pode-se realizar a etapa de teste, no qual o classificador recebe um conjunto de dados, necessariamente diferentes daqueles usado para treinamento, e os classifica com base no que foi ’aprendido’.

Entretanto, se o algoritmo não recebe dados rotulados na etapa de treinamento, o aprendizado é não-supervisionado. Geralmente, usam-se algoritmos de clusterização para identificar as classes dos dados, já que essas também não são providas ao algoritmo (HAN; KAMBER, 2006).

A forma mais comum de se treinar um classificador é fornecendo uma entrada, calculando uma saída e comparando o resultado com o rótulo correto. Mas, no caso do aprendizado por reforço, nenhum sinal da categoria, ou classe, correta é fornecido. Ao invés disso, o classificador recebe apenas um *feedback*, geralmente binário, indicando se o resultado está certo ou errado (DUDA; HART; STORK, 2000).

2.3.2 Algoritmos avaliados neste trabalho

Para este trabalho, foram utilizados três tipos diferentes de algoritmos de aprendizado de máquina supervisionado na etapa de classificação das opiniões. Foram eles: *Support Vector Machines* (SVM), Regressão Logística (RL) e o algoritmo de aprendizado profundo (*Deep Learning*).

2.3.2.1 Support Vector Machines (SVM)

Esse método foi desenvolvido por Cortes e Vapnik (1995) e usa hiperplanos como limites de decisão. O classificador foi desenvolvido para classificações binárias usando um hiperplano de separação ótimo entre as duas classes pela maximização da margem entre os pontos mais próximos de cada classe. Um hiperplano é uma superfície no espaço com várias dimensões, que é separado em duas metades de espaço. Uma vez que o SVM foi treinado, ele está apto a avaliar novas entradas relativas ao hiperplano divisor e classificá-las em uma categoria (READHEAD, 2012).

Dado um conjunto de treino com pares de dados anotados (x_i, y_i), onde $i = \{1, 2, 3, \ldots, l\}$, $x_i \in R^n$ e y_i em $\{1, -1\}^l$, o SVM precisa resolver o problema de otimização presente na Equa-

ção 2.1:

$$\begin{aligned} \min_{x,b,\xi} \quad & \frac{1}{2}W^T W + C\sum_{i=1}^{l}\xi_i \\ sujeito\ a \quad & y_i(W^T\phi(x_i)+b) \geq 1-\xi_i, \\ & \xi_i \geq 0, \end{aligned} \tag{2.1}$$

onde W é o parâmetro do peso atribuído às variáveis, ξ_i é a folga, ou correção de erro, adicionada e C é o fator de regularização (HSU et al., 2003). O objetivo do problema é minimizar $\frac{1}{2}W^T W + C\sum_{i=1}^{l}\xi_i$, onde o valor de $y_i(W^T\phi(X_i)+b)$ precisa ser maior ou igual a $1-\xi_i$ e o valor de ξ é bem pequeno.

A efetividade do classificador é definida pelo núcleo utilizado. O núcleo define a forma que os limites serão desenhados, por exemplo, usando um núcleo linear, como o SVC Linear utilizado neste trabalho, os limites serão todos linhas retas.

2.3.2.2 Regressão Logística (RL)

O modelo de Regressão Logística é um classificador linearizado probabilístico que usa um hiperplano para descrever o conjunto de dados. O classificador é parametrizado por uma matriz de peso W e um vetor de viés b. A classificação é realizada projetando um vetor de entrada em um conjunto de hiperplanos, cada um deles correspondendo a uma classe na base de dados. A distância de uma entrada até um hiperplano reflete a probabilidade dessa entrada ser um membro da classe correspondente (DEEPLEARNING.NET, 2017).

A probabilidade de um vetor de entrada x ser um membro da classe i, que, por sua vez, é um valor da variável estocástica Y, pode ser escrito como na Equação 2.2:

$$P(Y=i|x,W,b) = softmax_i(W_x+b) = \frac{e^{W_i x+bi}}{\sum_j e^{W_j}x+b_j}. \tag{2.2}$$

A saída do modelo, ou seja, a predição y_{pred} (DEEPLEARNING.NET, 2017), será a

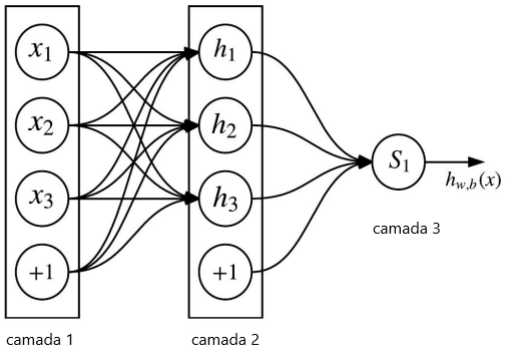
classe que obtiver a máxima probabilidade, como descrito na Equação 2.3:

$$y_{pred} = argmax_i P(Y = i|x, W, b) \tag{2.3}$$

2.3.2.3 Rede Neural (RN)

A rede neural *feedforwards* consiste em três camadas C1, C2 e C3, apresentado na Figura 2.4. A primeira camada é a de entrada(C1), que corresponde ao vetor de entrada (x_1,x_2,x_3) e intercepta o termo +1. A terceira camada representa a saída(C3), que corresponde ao vetor de saída (s_1).a segunda camada chama-se oculta(C3), cuja saída não é visível como uma saída de rede. Cada círculo em C1 representa um elemento no vetor de entrada, enquanto um círculo em C2 ou C3 representa um neurônio, o elemento básico de computação de uma rede neural. Também é chamada a C3 como função de ativação(h). Cada linha entre dois neurônios representa uma conexão para o fluxo de informação. Cada uma dessas conexões está associada a um peso, que é um valor que controla o sinal entre dois neurônios. A rede neural aprende a partir do resultado obtido e o ajuste feito aos pesos entre os neurônios com a informação que flui através deles. Os neurônios leem a saída da camada anterior, processam as informações e geram saída para os neurônios na próxima camada. Após o processo de treinamento, ele obterá uma forma complexa de hipóteses h, que se ajustam aos dados.

Figura 2.4 – Rede neural *feedforwards*



Fonte: Elaborado por Zhang, Wang e Liu (2018)

Entrando na camada oculta, percebe-se que cada neurônio de C2 tem como entrada x_1,x_2,x_3, com um valor constate +1 de C1, e o valor de saída é dada por $f(W^t x) = \sum_{i=1}^{3} W_i x_i + b$, e depois passa pela função de ativação. A funções de ativação mais comuns são sigmoid, função

tangente hiperbólica (tanh) ou função linear retificada (ReLU).

$$h(f(W^t x)) = sigmoid(W^t x) = \frac{1}{1 + \exp(-W^t x)} \tag{2.4}$$

$$h(f(W^t x)) = tanh(W^t x) = \frac{e^{W^t x} - e^{-W^t x}}{e^{W^t x} + e^{-W^t x}} \tag{2.5}$$

$$h(f(W^t x)) = ReLU(W^t x) = max(0, W^t x) \tag{2.6}$$

2.4 Aprendizado profundo (*Deep Learning*).

A aprendizado profundo por (WEHRMANN et al., 2017), é comumente associada às redes neurais profundas, tem ganho destaque na de análise de sentimentos e em outras tarefas da mineração de texto. As redes neurais convolucionais e as redes de memória de longo prazo têm sido amplamente empregadas na classificação de dados não estruturados, como é o caso de opiniões, assim como tem sido utilizada a incorporação de palavras para texto. O aprendizado profundo é conjunto de redes neurais em que aproxima ao trabalho feito pelo cérebro humano em que muito grande de conexões Deeplearning.net (2017)

2.4.1 Incorporação de palavra

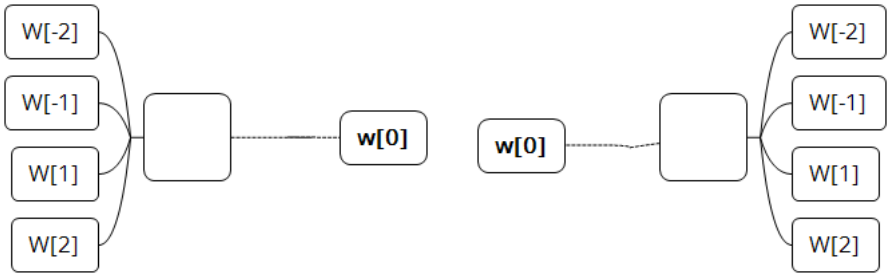
O vetores de palavras impacta indiretamente o quanto o classificador é considerado bom. O trabalho de Hartmann et al. (2017) constrói e analisa quatro estruturas para treino montadas que são: FastText, GloVe, Word2Vec e Wang2Vec sendo montada com 17 bases de dados de textos para português [2]. Para este trabalho, para este trabalho vai ser utilizadas o *Word2Vec* e *Wang2Vec* que é uma alteração do *Word2Vec*.

Vector space model O Word2Vec é um modelo de vetores de espaço(*vector space model*)

[2] disponível em: http://nilc.icmc.usp.br/embeddings

muito usado para processamento de linguagem natural (PLN) para a incorporação de palavras em que consiste em duas principais abordagens que são a Contínua Bag-of-Words(CBOW) e Skip-Gram mostrados na Figura 2.5. Para o CBOW é considerado, dado uma frase sem uma palavra, sendo assim realiza a busca das possíveis palavras para incorporação para palavra faltando. O Skip-Gram é o inverso, pois tem como entrada a palavra e retorna a possível frase. Para as duas abordagens o CBOW e Skip-Gram no modelo salvo caracterizam em uma matriz de pesos para cada palavra da dimensão, além da palavra. Em que o uso do Word2Vec resulta em um treino rápido de log-linear sendo capaz de obter informações semânticas em é proposto por Mikolov et al. (2013).

Figura 2.5 – Representação do CBOW e Skipgram



Fonte: Elaborado pelo autor

O Wang2vec é uma modificação do Word2Vec proposto por Ling et al. (2015), considerando a falta de ordem para a estrutura do texto original. Contendo duas diferenças: Para o CBOW a durante o processo de incorporação é feita com estrutura sintática das palavras no contexto original; e para o Skip-gram é dado para pesquisa de incorporação a a palavra e a posição onde está. Estando disponível para CBOW e o Skip-Gram.

2.5 Trabalhos relacionados

Para a busca de trabalhos relacionados, foi utilizado na *string* busca é ("*deep learning* "aprendizado profundo") ("*opinion mining* "mineração de opinião "*sentiment analysis* "analise de sentimento") ("*portuguese*") ("*political* "política"). Foram selecionado 5 trabalhos que apresentam aprendizado profundo, e serão explicados nas Seções a seguir.

2.5.1 Stance detection with bidirectional conditional encoding

O trabalho de Augenstein et al. (2016), é feita a detecção da posicionamento é um tarefa de classificar a atitude expressa em um texto em direção a um alvo pessoa "A"como “positiva”, “negativa” ou “neutra”. Nele foi experimentado uma codificação LSTM condicional, que cria uma representação dos *tweets* que depende do destino e demonstra que ele supera a codificação dos *tweets* e do destino de forma independente. O desempenho é melhorado quando o modelo condicional é aumentado com codificação bidirecional. A aplicação foi avaliada usando os dados de treino do SemEval 2016/ com o domínio de politica, utilizando técnica não-supervisionado de aprendizado. Obteve como resultado de 58,03% de f-medida.

2.5.2 Borrow a little from your rich cousin: using embeddings and polarities of english words for multilingual sentiment classification

o trabalho de Singhal e Bhattacharyya (2016) propõe o uso de aprendizado profundo com a utilização de diferentes linguas que são holandês, francês, espanhol, italiano, alemão, português e russo. Usando o *dataset* do Google News Corpus com revisões de filmes indianos, usando maquina de tradução para inglês e a incorporação de palavras também em inglês com o modelo Word2Vec de dimensão 300. Obteve o f-medida 79.9% com a rede convolucional para língua portuguesa.

2.5.3 Predicting polarities of tweets by composing word embeddings with long short-term memory

O trabalho de Wang et al. (2015) propõe uma reestruturação do processo de incorporação de palavras. Usando o corpus Stanford Twitter Sentiment junto com diversas técnicas de classificação, inclusive aprendizado profundo. Alcançado 87,20% para o f-medida para o aprendizado profundo.

2.5.4 A character-based convolutional neural network for language-agnostic Twitter sentiment analysis

O trabalho de Wehrmann et al. (2017) desenvolve uma classificação que usa para inglês, português, alemão, espanhol, juntas na mesma base e a redução das opiniões a *character* com uma modificação nos filtros. Alcançando 70.6% e 64.7% para acúracia e f-medida com a língua portuguesa.

No Quadro 2.2 seguir está relacionado os trabalhos relacionados e este. Com as características de cada um dos trabalhos, em que tem a língua usada, a quantidade de dados disponível no *dataset*, nome da base usada ou coletada, logo depois tem as técnicas de pre-processamentos usadas, em seguida tem os algoritmos de aprendizado utilizados e por fim tem as métricas usadas que são acurácia, precisão, revocação e o f-medida.

Quadro 2.2 – Visão geral dos trabalhos relacionados.

Trabalhos	Língua	Quantidade	Base	Técnica	Algorítimo	Acurácia	Precisão	Revocação	F-medida
Augenstein et al. (2016)	inglês	395,212	SemEval / Política	word2vec skip-gram	bidirecional	-	-	-	58.03%
Singhal e Bhattacharyya (2016)	multilíngue	6.826	Google Newsdataset	maquina tradutora	CNN	-	-	-	79.9%
Wang et al. (2015)	-	1,6 milhão	Stanford Twitter Sentiment corpus	word2vec Skipgram	SVM, MNB, MAXENT, MAX-TDNN, NBoW, DCNN, RAE, RNN, RNN, LSTM	-	-	-	87.20%
Wehrmann et al. (2017)	multilíngue	128.189	Twitter	N-gram, remoção de *stopwords* e stemmização	SVM, CNN e LSTM	70.60%	-	-	64.70%
Este trabalho	português	815	Votenaweb	word2vec, wang2vec e remoção de *stopwords*	rede profunda, SVM, Regressão linear	87.09%	87.17%	87.13%	87.07%

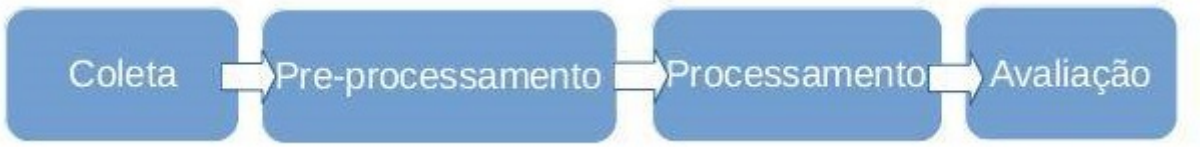
3 Método

Neste capítulo, é apresentado o método adotado . A Seção 3.1 apresenta a etapa de construção do Corpus. A Seção 3.2 apresenta as tarefas de pré-processamento realizadas para limpeza e estruturação do Corpus. As Seções 3.3 e 3.4 apresentam as técnicas utilizadas para processamento de validação da aplicação de Mineração de Opinião.

Neste estudo, foi realizada análise de sentimento a nível de documento, uma vez que são utilizadas bases de dados compostas por opiniões, que, apesar de curtos, podem conter mais de uma sentença.

A Figura 3.1 apresenta o método da aplicação de mineração de opinião utilizado. Ele é dividido em quatro etapas: “coleta dos dados”, na qual os comentários foram extraídos da página VOTENAWEB e anotados; “pré-processamento do texto”, na qual os comentários coletados foram estruturados antes de serem processados; “processamento”, na qual ocorreu a classificação dos comentários utilizando os algoritmos de aprendizado supervisionado; e “avaliação”, responsável pela obtenção dos resultados das classificações. Nas Seções subsequentes, cada uma das etapas é descrita com mais detalhes.

Figura 3.1 – Método adotado.



Fonte: Elaborado por Nascimento et al. (2017)

3.1 Coleta de dados

A coleta foi subdividida em três etapas descritas abaixo:

3.1.1 Extração das opiniões

A plataforma VOTENAWEB foi selecionada para a extração das opiniões, por seu volume de acessos. Como a mesma não disponibiliza API de acesso aos comentários, a extração foi realizada utilizando técnicas de raspagem de dados, também conhecida por *web scraping*. Foram coletadas 5 (cinco) das 15 (quinze) mil opiniões sobre o Projeto de Lei 1767-2015, que determina se o Rodeio e suas atividades relacionadas serão consideradas patrimônio cultural imaterial do Brasil. O projeto estava em tramitação entre os anos 2015 e 2016.

Assim como foi coletado do Impechment da ex-presidente Dilma Rousserff que determina o afastamento do cargo, contendo 14.631 comentários. Em que ambos têm a maior número de opiniões emitidas de acordo com a plataforma.

3.1.2 Anotação manual

Para construção da coleção dourada, utilizada nas etapas de treinamento e validação, 2 (dois) anotadores classificaram 500 (quinhentas) no trabalho de Nascimento et al. (2017) e para este trabalho foram acrescentado mais 4 anotadores divididos em duplas, nos quais anotaram 1000 rotulando as opiniões em uma das classes a seguir:

- Positivo: comentário contendo sentimento ou opinião positivas;
- Negativo: comentário contendo sentimento negativos ou opinião negativas;
- Ambos: quando o comentário continha, ao mesmo tempo, opiniões positivas e negativas;
- Neutro: quando o comentário não continha sentimento ou opinião;
- Descarte: propaganda ou anúncio;
- Ironia: opiniões que demonstram ironia através do uso de *emoticon* ou *hashtags*.

O Quadro 3.1 apresenta exemplos dos comentários anotados.

De acordo com Wiebe, Wilson e Cardie (2005) e Golden (2011), a capacidade humana para avaliação correta da subjetividade de um texto gira em torno de 72% a 85%, então cada opinião foi classificada por seis anotadores, sendo que para esse trabalho foram utilizado 4

Quadro 3.1 – Exemplo de classificação

Classes	Opiniões
Positivo	Parabéns capitão pelas palavras no congresso.
Negativo	Absurdo é pessoas que nunca viram de perto falarem coisas sem saber.
Ambos	Rodeio sim. Vaquejada não.
Neutro	conheçam antes de falar coisas que nunca viram
Descarte	Não esqueçam,agora dia 4 de outubro é dia nacional do rodeio.
Ironia	boxe, esgrima, MMA, isso pode?

Fonte: Elaborado por Nascimento et al. (2017)

anotadores dividindo em duplas para classificar mil opiniões como uma forma de minimizar classificações erradas. O Corpus utilizado foi criado neste trabalho Nascimento et al. (2017) e feita uma expansão, foi obtido o coeficiente de concordância Kappa (COHEN, 1960) no valor de 66,90%.

3.1.3 Adaptação

ParaPang e Lee (2008) discute que o tratamento de sarcasmo e ironia necessita de métodos mais sofisticados para a identificação.

Para realização os experimentos, foram selecionadas as classes "positiva", "negativa". Além disso, 2 (dois) data sets foram originados, conforme apresentado no Quadro 3.2. As classes desbalanceadas possuem classes com diferentes quantidades de opiniões.

Quadro 3.2 – Classes

Quantidade	Positivo	Negativo	Total
balanceada	306	306	612
desbalanceada	306	509	815

Fonte: Elaborado pelo autor(2019)

3.2 Pré-processamento

Para que os algoritmos de aprendizagem de máquina sejam utilizados, se faz necessário estruturar e limpeza dos textos. Neste trabalho, as seguintes técnicas de pré-processamento foram utilizadas: Tokenização, Remoção de *Stopwords*, Suavização e N-gram, enquanto que para aprendizado profundo se faz necessário da utilização da Tokenização, contador e a seleção de

palavras da incorporação de palavras. A biblioteca NLTK (Nature Language Toolkit) do Python foi utilizada nesta etapa.

3.2.1 Tokenização

A *tokenização* é responsável pela divisão do texto em palavras ou sentenças, chamadas de *tokens* (WEISS et al., 2005). O texto é dividido por um certo delimitador, em geral o carácter de espaço em branco. Na aplicação foi utilizado o *TweetTokenizer* da biblioteca NLTK. O Quadro 3.3 apresenta um exemplo.

Quadro 3.3 – Exemplo do Tokenizador

Entrada	[’como presente de natal a minha 3g está muito rápida obg @vivo’]
Saída	[’como’ , ’presente’ , ’de’ , ’natal’ , ’a’ , ’minha’ , ’3g’ , ’está’ , ’muito’ , ’rápida’ , ’obg’ , ’@vivo’]

Fonte: Elaborado por Nascimento et al. (2017)

3.2.2 *Bag-of-Word*

No Bag-of-Word, ou bolsa de palavras em português consiste em coletar todas as palavras que tem na base de dados e assim cada indice nas opiniões é trocado por 0 se existir a palavra e 1 se caso tiver, ou seja, todas as opiniões é tranfomordados em um grande lista de zeros e uns. Como mostrado no Quadro 3.4.

Quadro 3.4 – Exemplo do *Bag-of-Word*

Entrada	[’como’, ’presente’, ’de’, ’natal’, ’a’, ’minha’, ’3g’, ’está’, ’muito’, ’rápida’, ’obg’, ’@vivo’]
Saída	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Fonte: Elaborado pelo autor(2019)

3.2.3 Remoção de *stopwords* (RS)

Há palavras bastante comuns que têm pouco valor semântico para a identificação de um texto e que podem ser retiradas sem danos. Tais palavras são chamadas de *stopwords* (MAN-

NING; RAGHAVAN; SCHÜTZE, 2009). Em alguns casos, a retirada das *stopwords* melhora a classificação de um texto mas, em alguns casos pode piorar (TELES; SANTOS; SOUZA, 2016). A tabela 3.5 apresenta um exemplo de opinião com a remoção das *stopwords*.

Quadro 3.5 – Exemplo de remoção de *stopwords*

Entrada	[’já’ , ’que’ , ’animal’ , ’nenhum’ , ’é’ , ’pra’ , ’culturalizar’ , ’ou’ , ’pra’ , ’estar’ , ’fazendo’ , ’torneios’ ,’com’,’ele’ , ’int’ , ’acaba’ , ’com’ , ’os’ , ’centro’ , ’de’ , ’macumba’ , ’que’ , ’usam’ , ’:’ , ’galinha’ ,’, bode’ , ’ , gato’ ,’...’ ,’e’ , ’ninguém’ , ’como’ , ’frango’ , ’ , galinha’ , ’ , , galeto’ , ’ , carne’ , ’ , peru’ , ’no’ ,’natal’ , ’acabou’ , ’...’]
Saída	[’animal’ , ’nenhum’ , ’é’ , ’pra’ , ’culturalizar’ , ’pra’ , ’estar’ , ’fazendo’ , ’torneios’ , ’int’ , ’acaba’ , ’centro’ ,’macumba’ , ’usam’ , ’:’ , ’galinha’ , ’ , , bode’ , ’ , gato’ , ’...’ ,’ninguém’ , ’frango’ , ’ , galinha’ , ’ , galeto’ ,’ , carne’ , ’, peru’ , ’natal’ , ’acabou’ , ’...’]

Fonte: Elaborado por Nascimento et al. (2017)

3.2.4 TF-IDF

A medida TF-IDF são dois termos: o primeiro calcula a Frequência de Termo normalizada (TF), em que é o número de vezes que uma palavra aparece em um documento, dividida pelo número total de palavras nesse documento; o segundo termo é a Frequência de Documento Inversa (IDF), calculado como o logaritmo do número de documentos no corpus dividido pelo número de documentos em que o termo específico aparece.

$$TFIDF = \frac{\frac{FPD}{TPD}}{\log \frac{TDC}{TDPC}} \tag{3.1}$$

3.2.5 Incorporação de palavras

Para incorporação de palavra, chamadas de Word2vec e Wang2vec foram empregadas abordagens previamente treinadas no trabalho de Hartmann et al. (2017) com o modelo de representação Skip-gram com dimensão igual a 50. Em que associa cada palavra a um vetor de pesos para incorporação

3.3 Processamento

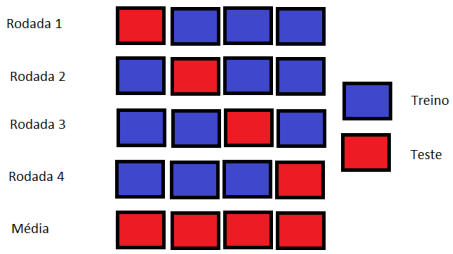
No processamento, é realizada a efetivação da classificação da polaridade de opiniões.Para esta etapa, foi utilizada a Scikit-Learn, uma biblioteca de código aberto com os algoritmos de aprendizado de máquina da linguagem de programação Python. Com a Scikit-Learn, foram avaliados os seguintes algoritmos: o algoritmo SVC Linear do SVM, e o algoritmo de Regressão Logística com o modelo Linear, semelhando ao trabalho de castro2017smoothed que obteve resultados ótimos.

A biblioteca Tensorflow estruturaliza as camadas de cada parte da aprendizagem de maquina e Keras que é uma API de alto nível para construir e fazer treinamentos de modelos de aprendizado profunda, ambas estão disponíveis no Python. São usadas para prototipagem rápida, e pesquisa avançada.

3.4 Avaliação

Para a etapa de avalização, foi escolhido o método de validação cruzada K-fold (com k igual a 4) representado na Figura 3.2.

Figura 3.2 – Validação K-fold.



Fonte: Elaborado pelo autor(2019)

Este método consiste em dividir a base em quatro partes de mesmo tamanho ou próximo, com 3 delas sendo destinadas para o treinamento, e a outra parte para teste do classificador. Por fim, uma média dos resultados é obtida, expressa utilizando as medidas de acurácia, precisão, revocação e f-medida.

4 Resultados

Neste capítulo são explicadas as configurações, juntamente com seus respectivos resultados alcançados. Também são apresentadas, na Seção 4.1.2, a discussão e a avaliação dos classificadores SVM, Regressão Logística e o aprendizado profundo.

4.1 Configuração do experimento

Na etapa de pré-processamento, foram usados os dois algoritmos de aprendizagem de maquina que obtiveram melhores resultados para análise de opinião de cidadãos que são o SVC Linear e a Regressão Linear, conforme apresentado no trabalho de (NASCIMENTO et al., 2017).Utilizado configuração do trabalho com o uso da tokenização, a remoção de *stopwords* e remoção de acentuação. Com o aprendizado profundo foi utilizada a tokenização, o *Bag-of-Words* e variando a técnica de incorporação de palavra. Para aprendizado de máquina e o aprendizado profundo foram divididas em duas bases de experimentos. Para melhor entendimento para os experimentos de aprendizado de maquina está representado no Quadro 4.1 e o aprendizado profundo no Quadro 4.2.

Quadro 4.1 – Configuração experimento do aprendizado de máquina.

Número	Classe	Técnica
1	balanceada	remoção de *stopwords*, remoção de acentuação
2	desbalanceada	remoção de *stopwords*, remoção de acentuação

Fonte: Elaborado pelo autor(2019)

Quadro 4.2 – Configuração experimento do aprendizado profundo.

Número	balanceada	desbalanceada	Wang2Vec	Word2Vec
1	x		x	
2	x			x
3		x	x	
4		x		x

Fonte: Elaborado pelo autor(2019)

A configuração do aprendizado profundo está representada no Quadro 4.3 a primeira camada de incorporação com os pesos, seguindo pela uma camada de operação de pool média global para deixar linearizar para não ocasionar aprendizado muito específico, um Flatten,

seguindo por 3 camadas densas de rede neural com funções de ativação de *softmax*, retificador, tangente hiperbólica,cada uma respectivamente .

Quadro 4.3 – Estrutura do aprendizado profundo.

	Camada	entrada	saída
Camada 1	entrada		3014
Camada 2	embedding	3014	50
Camada 3	globalAveragePooling1D	50	50
Camada 4	flatten	50	50
Camada 5	dense	50	25
Camada 6	dense	25	12
Camada 7	dense	12	1
Camada 8	saída	1	

Fonte: Elaborado pelo autor(2019)

Para o otimizador, foi utilizado o *AdamOptimizer* disponibilizado na biblioteca *TensorFlow*, para o calculo da taxa de erro foi utilizado a função *binary crossentropy*. Para o treino foram adotadas 30 repetições para quantidade de épocas, com o tamanho do *batch* de 1.

4.1.1 Resultados do experimentos

As Tabelas 4.1 tem os resultados alcançados para a base balanceada por classes e 4.2 tem para a base desbalanceadas e 4.3 apresentam os resultado das métricas com a media aritmética de cada uma. A primeira tabela apresenta os resultados por classes, para a positiva e negativa respectivamente em cada uma das métricas e algoritmo. Na primeira primeira tabela experimento contendo apenas classes balanceadas na coleção dourada, e o segunda parte com as classes desbalanceada. Em que destaca-se a rede profunda que obteve 100% para a precisão para classe positiva.

Na segunda tabela é apresentado a acurácia e a media aritmética da precisão, revocação, e o f-medida das classes. Na Tabela 4.3 podemos destacar o classificador Regressão Linear com o melhor resultado para as quatro métricas, com resultados maiores que os demais para o cenário das classes desbalanceadas e das balanceadas.

Tabela 4.1 – Resultados obtidos por classe balanceadas.

Algoritmo	Precisão		Revocação		F-medida	
	Positiva	Negativa	Positiva	Negativa	Positiva	Negativa
RL	85.91%	87.80%	85.35%	88.90%	86.83%	87.31%
SVC Linear	83.33%	86.41%	83.72%	86.29%	84.74%	85.14%
RP - Wang2Vec	100.0%	60.50%	41.97%	100.0%	59.13%	75.39%
RP - Word2vec	92.85%	80.72%	80.24%	93.05%	86.09%	86.45%

Fonte: Elaborado pelo autor(2019)

Tabela 4.2 – Resultados obtidos por classe desbalanceadas.

Algoritmo	Precisão		Revocação		F-medida	
	Positiva	Negativa	Positiva	Negativa	Positiva	Negativa
RL	98.36%	82.75%	75.95%	95.73%	82.73%	91.04%
SVC Linear	95.38%	83.68%	78.56%	92.79%	82.17%	90.16%
RP-Wang2Vec	100.0%	74.40%	45.56%	100.0%	62.60%	85.32%
RP-Wang2Vec	100.0%	69.83%	31.64%	100.0%	48.07%	82.23%

Fonte: Elaborado pelo autor(2019)

Tabela 4.3 – Resultados obtidos.

Configuração	classificador	Acúracia	Precisão	Revocação	F-medida
Balanceada	Regressão Linear	**87.09%**	**87.17%**	**87.13%**	**87.07%**
	SVC Linear	84.96%	85.07%	83.89%	84.94%
	RP - Wang2Vec	69.28%	80.25%	70.98%	67.26%
	RP - Word2Vec	86.27%	86.79%	86.65%	86.27%
Desbalanceada	Regressão Linear	**88.22%**	**89.25%**	85.84%	**86.89%**
	SVC Linear	87.36%	87.69%	**85.97%**	86.17%
	RP - Wang2Vec	78.92%	87.20%	72.78%	73.9%
	RP - Word2Vec	73.52%	84.91%	65.82%	65.15%

Fonte: Elaborado pelo autor(2019)

4.1.2 Discussão

A Tabela 4.3 apresenta os resultados das métricas do experimento contendo o mesmo numero de classes na coleção dourada, dos resultados obtidos em todas elas a regressão obteve melhores resultados. Para a RP obteve 86,27%, enquanto que a RL obteve 87,09% e o SVM obteve 84,96% para a acurácia, ou seja, não tão distantes dos outros dois classificadores, é possível afirmar que com o Wang2Vec está acertado mais uma classe que a outra devido relativa diferença entre a precisão e a revocação, para a Word2Vec tem quase insignificante diferença, ou seja, está mantendo uma bom equilíbrio de acerto por classes como mostrado na Tabela 4.1.

No experimento para coleção dourada com classes desbalanceadas mostrada na Tabela 4.3, manteve-se melhor a Regressão como melhor resultado com 88,22% não tão distante da rede profunda com 78,92% e para as demais métricas manteve muito próximo os valores,

enquanto que para a rede profunda apontou uma diferencia significante para a revocação e o f-medida. A rede profunda alcançou uma taxa de precisão da classe positiva de 100%, enquanto que caiu o da classe negativa como mostrado na Tabela 4.2.

Do ponto de vista que a Tabela 4.3 mostrando dois cenários, no qual o primeiro com classes balanceadas tem menos dados de treino, ou seja, consequentemente menos de teste, é possível cria a hipótese de que a rede profunda obtêm melhores resultados para bases que tenham a mesma quantidade de elementos de treino por classe.

Para o trabalho de Nascimento et al. (2017), tendo como foco o uso de aprendizagem de maquina para mineração de opinião alcançando 81,26%, assim como não tendo como enfase o aprendizado profundo, que propõe resultados melhores ainda ficando próximo dos resultados com uma diferença percentual de 5,01% para rede profunda neste trabalho. Também é importante resaltar que quantidade de dados usada como técnicas de pre-processamento e processamento.

Em Singhal e Bhattacharyya (2016) e Wehrmann et al. (2017) por terem objetivo multilíngue ainda assim alcançaram resultados satisfatórios para a aplicação de mineração de opiniões em português com bases de dados maiores que foi de 64.70% conduto ainda neste trabalho obteve um f-medida um pouco melhor, com 84,90%.Com exceção de Nascimento et al. (2017), foram utilizadas bases grandes para o restantes.

5 Conclusão

Neste capítulo, são apresentados as conclusão da realização, proposta de trabalhos futuros 5.1 e limitações e ameaças 5.2 deste Projeto de Conclusão de Curso.

Neste trabalho foi proposto analise de técnicas de aprendizado supervisionado em que mostrou o aprendizado profundo é possível para mineração de opinião, devido a incorporação de palavras com a abordagem de incorporação de palavras que possibilitou melhores resultados. Assim como o aprendizado de maquina ainda mostrou-se relevante para mineração de opinião. Tanto o aprendizado profundo como o de maquina mostram impactantes para a classificação de opinião de plataformas de participação eletrônica.

Assim como proposto foi expandido aqui o corpus do trabalho de Nascimento et al. (2017) que contem 319 opiniões para 815 opiniões.

Aplicou-se diferentes técnicas de pre-processamento, com algoritmos de aprendizagem avaliando-os a partir das métricas.

Em virtude dos fatos mencionados mostrou-se que o tratamento dos dados, a utilização de aprendizado de maquina e aprendizado profundo são possíveis para mineração de opinião de cidadãos.

Um artigo completo contendo as parte configurações propostas e os resultados apresentados neste trabalho foi submetido ao *VI Simpósio Brasileiro de Tecnologia da Informação* (SBTI 2017) e encontra-se publicado.

5.1 Proposta para trabalhos futuros

Foi possível utilizar uma técnica chamada Wang2Vec, contudo em Hartmann et al. (2017) montou vários modelos, dimensões, técnicas para este trabalho foi melhor com 50 dimensões. Um possível trabalho futuro seria a utilização de uma base de treino maior, assim podendo analisar com ênfase nessas variações para mineração de opinião. Assim como utiliza-las em aprendizado de maquina tanto profundo por obter um ganho significativo no desempenho.

Identificar variações de polaridade a partir do tempo,para quais aspectos estão em evidencia e as polaridades que estão influenciando esse aspecto.

Uma areá interessante para pesquisa é de analise de sentimento aplicado a importância dando pelos gestores públicos a partir de suas ações no meio publico tanto em plataformas de participação e colaboração eletrônica como as redes sociais. Assim como velocidade que um aspecto se propagar nas redes sociais.

5.2 Limitações e desafios

A primeira limitação foi que não tratar o sarcasmo e ironia, mesmo que na etapa de construção do corpus tenha rotulado e retirado da coleção dourada. A identificação tem um grau de dificuldade para o entendimento para seres humano elevado, até para pessoas próximas a quem emitiu a opinião pode entender inadequadamente. Para o aprendizado supervisionado precisa dessas opiniões corretamente rotulados para assim seja aprendido.

A segunda é a quantidade de informações para rede neural poder aprender, além de recursos de hardware CPUs,GPUs. Assim consequentemente para aprendizado profundo mesmo tendo o recurso do Google disponível que foi especificamente desenvolvido para melhorar o desempenho e redução de gastos financeiros que é a TPU.

A terceira é que as configurações do aprendizado profundo não mantenha seu desempenho para uma base maior com as mesmas configurações, assim como para a regressão e o SVM podendo não adequasse. Tanto no tamanho do vocabulário quanto o domínio da base.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, A. L. F. et al. In: *Anais do 20º Brazilian Symposium on Multimedia and the Web*. Nova York, NY, EUA: ACM, 2014. (WebMedia '14), p. 123—130.

AUGENSTEIN, I. et al. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*, 2016.

BALAZS, J. A.; VELÁSQUEZ, J. D. Opinion Mining and Information Fusion: A survey. *Information Fusion*, v. 27, p. 95–110, 2016.

BRESSAN, R. T. U. Dilemas da rede: Web 2.0, conceitos, tecnologias e modificações. *Intercom*, p. 1–13, 2007.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. *Journal of machine learning research*, v. 12, n. Aug, p. 2493–2537, 2011.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: *Anais da 12ª International Conference on World Wide Web*. Nova York, NY, EUA: ACM, 2003. (WWW '03), p. 519–528.

DEEPLEARNING.NET. *Classifying MNIST digits using Logistic Regression*. 2017. Disponível em: <http://deeplearning.net/tutorial/logreg.html>. Acesso em: maio de 2017.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*. 2. ed. [S.l.]: Wiley-Interscience, 2000.

FELDMAN, R. Techniques and applications for sentiment analysis. *Commun. ACM*, ACM, Nova York, NY, EUA, v. 56, n. 4, p. 82–89, abr. 2013. ISSN 0001-0782.

FRANKLIN, A.; EBDON, C. Aligning priorities in local budgeting processes. *Journal of Public Budgeting, Accounting & Financial Management*, PrAcademics Press, Florida Atlantic University, v. 16, n. 2, p. 210, 2004.

FREITAS, J. L. d. et al. Participação eletrônica, transparência e accountability no gabinete digital sob a lente da teoria da ação comunicativa. Pontifícia Universidade Católica do Rio Grande do Sul, 2015.

FUNG, A. Varieties of participation in complex governance. *Public administration review*, Wiley Online Library, v. 66, n. s1, p. 66–75, 2006.

GOLDEN, P. *Write here, write now*. 2011. Disponível em: <https://www.research-live.com/article/features/write-here-write-now/id/4005303>. Acesso em: maio de 2017.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. 2. ed. São Francisco, CA, EUA: Morgan Kaufmann, 2006.

HARTMANN, N. et al. Portuguese word embeddings: evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

HSU, C.-W. et al. A practical guide to support vector classification. Taipei, Taiwan, 2003.

IRVIN, R. A.; STANSBURY, J. Citizen participation in decision making: Is it worth the effort? *Public administration review*, Wiley Online Library, v. 64, n. 1, p. 55–65, 2004.

KING, C. S.; FELTEY, K. M.; SUSEL, B. O. The question of participation: Toward authentic public participation in public administration. *Public administration review*, JSTOR, p. 317–326, 1998.

LING, W. et al. Two/too simple adaptations of word2vec for syntax problems. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2015. p. 1299–1304.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: ____. *Mining Text Data*. Boston, MA, EUA: Springer US, 2012. p. 415–463. ISBN 978-1-4614-3223-4.

MACINTOSH, A. Characterizing e-participation in policy-making. In: IEEE. *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. [S.l.], 2004. p. 10–pp.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to information retrieval/Christopher D*. [S.l.]: Cambridge University Press, Cambridge, England, 2009.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, v. 5, n. 4, p. 1093 – 1113, 2014.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

MITCHELL, T. M. *Machine Learning*. 1. ed. [S.l.]: McGraw-Hill, 1997. (McGraw-Hill International Editions).

NASCIMENTO, E. et al. Análise de opinião dos usuários de plataformas de e-participação e e-colaboração: um estudo de caso do portal votenaweb. In: *VI Simpósio Brasileiro de Tecnologia da Informação*. [S.l.: s.n.], 2017. p. 144—159.

NELSON, N.; WRIGHT, S. et al. *Power and participatory development: Theory and practice*. [S.l.]: ITDG Publishing, 1995.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, v. 2, n. 1-2, p. 1–135, 2008.

RAVI, K.; RAVI, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, v. 89, p. 14–46, 2015.

READHEAD, J. *Machine Learning: How Support Vector Machines can be used in Trading*. 2012. Disponível em: <https://www.mql5.com/en/articles/584>.

SANTOS, J. *Funcionalidades, características e limitações das plataformas de participação e colaboração: uma análise comparativa.* 2017.

SANTOS, J. A. P. dos; NETO, J. d. S. C.; SOUZA, E. Funcionalidades, características e limitações das plataformas de participação e colaboração: Uma análise comparativa-functionalities, characteristics, and limitations of participation platforms and collaboration: A comparative analysis. *GESTÃO. Org-Revista Eletrônica de Gestão Organizacional-ISSN: 1679-1827*, v. 15, n. 2, 2017.

SINGHAL, P.; BHATTACHARYYA, P. Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. [S.l.: s.n.], 2016. p. 3053–3062.

SOUZA, E. et al. Characterizing opinion mining: A systematic mapping study of the portuguese language. In: ____. *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings*. Cham: Springer International Publishing, 2016. p. 122–127. ISBN 978-3-319-41552-9.

TELES, V.; SANTOS, D.; SOUZA, E. Uma análise comparativa de técnicas supervisionadas para mineraçao de opiniao de consumidores brasileiros no twitter. *Proceedings of the XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2016)*, p. 217–228, 2016.

VITÓRIO, D. et al. Investigating opinion mining through language varieties: a case study of Brazilian and European Portuguese tweets. In: *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*. Uberlândia, Brazil: Sociedade Brasileira de Computação, 2017. p. 43–52. Disponível em: <https://www.aclweb.org/anthology/W17-6607>.

WANG, X. et al. Predicting polarities of tweets by composing word embeddings with long short-term memory. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. [S.l.: s.n.], 2015. v. 1, p. 1343–1353.

WEHRMANN, J. et al. A character-based convolutional neural network for language-agnostic twitter sentiment analysis. In: IEEE. *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2017. p. 2384–2391.

WEISS, S. M. et al. *Text mining: predictive methods for analyzing unstructured information*. [S.l.]: WorldCist'16, 2005.

WIEBE, J.; WILSON, T.; CARDIE, C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, v. 39, n. 2, p. 165–210, 2005.

ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1253, 2018.