UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

RAFAEL DUTRA CAVALCANTI

CLASSIFICAÇÃO DE TENDÊNCIAS POLÍTICAS EM NOTÍCIAS VIA MINERAÇÃO DE TEXTO E REDES NEURAIS SEM PESO

Rio de Janeiro
2017

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

RAFAEL DUTRA CAVALCANTI

CLASSIFICAÇÃO DE TENDÊNCIAS POLÍTICAS EM NOTÍCIAS VIA MINERAÇÃO DE TEXTO E REDES NEURAIS SEM PESO

Dissertação de Mestrado submetida ao Corpo Docente do Departamento de Ciência da Computação do Instituto de Matemática, e Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Orientador: Daniel Sadoc Menasché
Co-orientador: Priscila Machado Vieira Lima

Rio de Janeiro
2017

CBIB Cavalcanti, Rafael Dutra

Classificação de Tendências Políticas em Notícias via Mineração de Texto e Redes Neurais Sem Peso / Rafael Dutra Cavalcanti. – 2017.
81 f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós-Graduação em Informática, Rio de Janeiro, 2017.

Orientador: Daniel Sadoc Menasché.
Co-orientador: Priscila Machado Vieira Lima.

1. Mineração de Textos. 2. Redes Neurais sem Peso. 3. Descoberta de Conhecimento em Dados não Estruturados. – Teses. I. Menasché, Daniel Sadoc (Orient.). II. Lima, Priscila Machado Vieira (Co-orient.). III. Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Programa de Pós-Graduação em Informática. IV. Título

CDD

RAFAEL DUTRA CAVALCANTI

Classificação de Tendências Políticas em Notícias via Mineração de Texto e Redes Neurais Sem Peso

Dissertação de Mestrado submetida ao Corpo Docente do Departamento de Ciência da Computação do Instituto de Matemática, e Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Aprovado em: Rio de Janeiro, 6 de OUTUBRO de 2017.

Prof. Dr. Daniel Sadoc Menasché (Orientador)

Profa. Dra. Priscila Machado Vieira Lima (Co-orientador)

Prof. Dr. Felipe Maia Galvão França

Profa. Dra. Jonice de Oliveira Sampaio

Prof. Dr. João Carlos Pereira da Silva

Dedico este trabalho à minha esposa Anne Caroline e ao nosso filho Davi.

AGRADECIMENTOS

Agradeço à minha esposa, Anne Caroline, por entender minha ausência em diversos momentos por conta dos estudos e estar ao meu lado em todas as decisões e dificuldades enfrentadas durante o curso, sempre buscando fazer com que houvesse o melhor ambiente para que eu pudesse realizar os estudos. Mãe dedicada, se absteve de seu lazer por diversas vezes e em muitas outras, deixou de dormir para cuidar do nosso filho Davi, para que eu pudesse ter uma boa noite de sono.

Agradeço aos meus pais, Cely Dutra e Roberto Cavalcanti, por estarem sempre ao meu lado, demonstrando preocupação com o andamento dos estudos, constantemente enviando vibrações positivas e se dispondo para ajudar em qualquer necessidade que pudesse ter. Agradeço à minha irmã Ana Paula Dutra, ao meu sogro Ricardo Maia e à minha sogra Ana Cristina Maia pelas boas energias enviadas e palavras de motivação proferidas em diversos momentos.

Agradeço ao Coronel Eraldo dos Santos Filho pelas orientações e incentivo. Agradeço ao Capitão-de-Corveta André Paim Gonçalves, pelas orientações acerca do curso e pela disponibilidade para diversos esclarecimentos. Também agradeço ao Capitão-de-Corveta (EN) Vilc Queupe Rufino pelas dúvidas esclarecidas e pelas horas de estudos compartilhadas.

Agradeço aos Professores Felipe França, Jonice Sampaio e João Carlos pela disponibilidade e orientações. Agradeço aos orientadores Daniel Sadoc e Priscila Lima pelos conhecimentos técnicos passados e pelo incentivo ao tema deste trabalho, que serão de grande valia na minha carreira.

Agradeço aos colegas Fabio Rangel e Fabrício Firmino pelos esclarecimentos acerca de suas implementações da WiSARD e outros esclarecimentos que foram muito úteis no desenvolvimento deste trabalho. Sou grato ao colega Douglas Cardoso pela disponibilidade para esclarecimentos de sua implementação da ClusWiSARD. Também agradeço ao Massimo De Gregorio pelas orientações e materiais relacionados à WiSARD cedidos.

Por fim, agradeço à Marinha do Brasil pela oportunidade de realizar um curso desta magnitude em uma Universidade conhecida mundialmente. A conclusão do mestrado é um motivo de grande satisfação e gera uma grande ansiedade com relação à utilização dos conhecimentos técnicos adquiridos.

RESUMO

Uma notícia tendenciosa é, às vezes, bem suave para o interlocutor, e alcança seu objetivo de influenciar a opinião do leitor no mesmo sentido. Nos dias atuais, devido a quantidade de informações existentes, muitas pessoas sentem dificuldades em avaliar a ideia principal do conteúdo de uma notícia ou se existe alguma tendência, no caso deste trabalho, política.

Nesta dissertação, buscamos a identificação de polaridade em notícias políticas em português através do processo de mineração de dados textuais com a utilização da Rede Neural sem Peso WiSARD e de uma derivação, a ClusWiSARD. O WiSARD funciona através de uma estrutura de discriminadores, onde cada discriminador é responsável por identificar uma classe. Realizamos avaliações relacionadas ao corpo da notícia e à manchete da notícia e realizamos uma avaliação de um veículo de mídia amplamente conhecido. Obtivemos acurácia de cerca de 90% ao utilizar o corpo da notícia completo e acurácia de cerca de 75% ao considerar apenas manchetes. Além disso, também fazemos uma análise temporal sobre a dinâmica política das tendências.

Palavras-chave: Mineração de Textos, Redes Neurais sem Peso, Descoberta de Conhecimento em Dados não Estruturados.

ABSTRACT

Biased news can influence the reader's opinion in subtle ways. Nowadays, due to the unprecedented amount of information created and made available through social media, the identification of biases is increasingly challenging. In the domain of politics, addressing the challenge is particularly relevant.

In this dissertation, we seek the identification of polarity in Portuguese political news through the process of textual data mining using the WiSARD Weightless Neural Network, and one of its extensions, the ClusWiSARD. The WiSARD classifier works through a structure of discriminators, where each discriminator is responsible for identifying a class. We assessed polarity using the body and the headline of news published in widely known media vehicles. The obtained results are encouraging, indicating the feasibility of automatic and efficient bias detection. We obtained accuracy of about 90% when using full body news and accuracy of the 75% when considering only headlines. In addition, we also perform a temporal analysis on the political dynamics of bias.

Keywords: Text Mining, Weightless Neural Networks, Discovery of Knowledge in Non-Structured Data.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento tecnológico dos computadores e a popularização dos dispositivos digitais móveis provocaram um aumento gigantesco no número de documentos digitais existentes. Diante do exponencial crescimento da quantidade de dados digitais em formato textual gerados pelas mais diversas fontes como instituições governamentais, militares, civis, empresas e usuários comuns que trafegam pela internet, identificar a categoria de cada documento por meio de sua leitura torna-se um desafio.

A mineração de textos (ou *text mining*) é um campo relacionado com diversas disciplinas como por exemplo aprendizado de máquina, recuperação da informação, processamento de linguagem natural e estatística. Com a mineração de texto, é possível obter informações e chegar a conclusões sobre um conjunto de documentos, através da análise dos dados.

A rede neural sem peso WiSARD é um classificador utilizado para o reconhecimento de padrões e utiliza valores binários em suas entradas. A WiSARD pode realizar a classificação de acordo com os padrões apresentados previamente durante a fase de treinamento. A classe de cada padrão apresentado à rede é associada a uma estrutura denominada discriminador. Cada discriminador é constituído de memórias RAM, que realiza o armazenamento do conhecimento por meio de tabelas-verdade.

Explorando o processo de mineração de textos e a WiSARD, buscamos classificar automaticamente as notícias com base em sua polaridade política. Em particular, nos concentramos em notícias políticas recentes do Brasil, coletadas das páginas dos partidos políticos que consideramos possuírem posições ideológicas opostas.

Como a identificação de polaridade é subjetiva, para ter uma base sólida, as fontes de dados selecionadas foram os *feeds* de notícias dos sítios dos partidos: Partido do Movimento Democrático Brasileiro (PMDB),[1] Partido da Social Democracia Brasileira (PSDB),[2] Partido dos Trabalhadores (PT)[3] e o Partido Socialismo e Liberdade (PSOL).[4]

Levamos em consideração neste trabalho que as fontes de notícias utilizadas gerem *feeds* de notícias com sua respectiva opinião implícita. Nossos resultados de classificação explicam o estilo de escrita dos conjuntos de autores e dos grupos por posição ideológica, juntamente com as diferenças no vocabulário que normalmente usam. Portanto, enquadramos o problema de identificação de polarização como um problema de reconhecimento de origem do texto.

Nosso problema consiste em identificar a origem e a polaridade, para cada um dos artigos no banco de dados selecionado. Abordamos algumas questões. A primeira é a viabilidade de se classificar automaticamente notícias políticas. A segunda refere-se às vantagens e desvantagens do classificador WiSARD [3] em relação à precisão e eficiência/desempenho, comparada aos classificadores SVM [53], Regressão Logística [29], *Naive Bayes* [33] e *Gradient Tree Boosting* [36]. A terceira é a utilização da ClusWiSARD [14] na avaliação de um noticiário, por meio de aprendizado semi-supervisionado.

Nós obtivemos uma resposta afirmativa à primeira pergunta e identificamos o WiSARD como uma ferramenta simples e eficiente para realizar a classificação. Portanto, identificar a viabilidade de classificar as fontes e as polaridades apenas com base no conteúdo do texto, com acurácia de cerca de 90%, é a nossa maior

[1]Site do PMDB: `http://www.pmdb.org.br`
[2]Site do PSDB: `http://www.psdb.org.br`
[3]Site do PT: `http://www.pt.org.br`
[4]Site do PSOL: `http://www.psol50.org.br`

contribuição. Como segunda contribuição, mostramos que mesmo arquiteturas bem simples, como o WiSARD e a ClusWiSARD, são suficientes para eficientemente executar a tarefa de classificação.

1.1 CONSIDERAÇÕES SOBRE AS POSIÇÕES PARTIDÁRIAS

Identificar a posição ideológica de um partido político é uma atividade complexa, ainda não possui um consenso e proporciona muitos debates por cientistas políticos, sociólogos e pesquisadores [54]. O trabalho de [51] apresenta essa dificuldade ao relacionar coligações partidárias, fatos noticiados, pesquisas de preferência de candidatos à modificação de opinão pública. Para exemplificar uma mudança de posicionamento partidário, podemos citar um fato que ocorreu na Dinamarca. Um partido denominado Esquerda Radical (*Radikale Venstre*) depois de variações de posicionamentos durante anos, hoje assume uma postura como partido de centro [52, 58].

Com o objetivo de realizar uma divisão das notícias de acordo posição ideológica dos partidos políticos, buscamos agrupar a notícias de acordo com o contexto histórico dos partidos avaliados. Consideramos que o PMDB e o PSDB possuem afinidades, de acordo com [12, 19, 59]. Com outra posição partidária, consideramos que os partidos PT e PSOL possuem uma afinidade em sua linha ideológica, conforme pode ser observado em [56, 55, 50]. Para fins de classificação por polaridade política, consideramos que as notícias podem pertencer a uma de duas classes, `PMDB/PSDB` e `PT/PSOL`.

1.2 MOTIVAÇÃO

Nos dia atuais, o processo de tomada de decisão de grande parte das instituições são conduzidos de acordo com informações extraídas dos dados [31]. Grande parte dos dados das instituições se encontram em formato não estruturado, seja por meio de *e-mails*, manuais, memorandos, relatórios, projetos [17]. Assim como nas instituições, a *web* também possui dados não estruturados em sua maior parcela. Na *web*, os dados são produzidos principalmente pelas redes sociais, sites de notícias e páginas sobre os mais diversos assuntos. O quadro apresentado é o fato motivador deste trabalho, que foca no processo de mineração de texto, com a finalidade de obter informação a partir de textos escritos em português do Brasil. Este trabalho se restringirá à classificação de notícias políticas.

O modelo individual para polaridade se expandiu para a sociedade moderna, que hoje em dia é indiscutivelmente mais polarizada em diversos assuntos. Uma recente declaração do papa faz referência a um "vírus da polarização" [46]. Pouco tempo após a declaração, o presidente dos Estados Unidos, Barack Obama, discursou sobre as causas de tal recente aumento da polarização [34]. Em seu discurso, ele apontou a ampla publicação de artigos tendenciosos nos meios de comunicação e redes sociais como um dos principais motores da polarização. A identificação de opiniões disfarçadas nas notícias é um desafio importante a ser enfrentado na busca de uma sociedade mais transparente e harmônica.

Conforme é constatado por [5], onde são estudados os tópicos de notícias que têm maior atenção da mídia e seus padrões de evolução temporal, a quantidade de notícias que cobrem um determinado assunto aumenta e diminui devido a muitos motivos. Também é apresentada a importância de se entender corretamente a dinâmica da cobertura das notícias, pois o período em que uma notícia é veiculada consegue articular as percepções dos leitores sobre a importância do problema.

Considerando este contexto, até mesmo os políticos, podem ser influenciados quando priorizam a opinião pública para a tomada de decisões.

O momento político do Brasil proporcionou uma imensa quantidade de notícias e informações diversas acerca de partidos políticos a favor da estrutura governamental atual e partidos políticos de oposição ao governo. Os noticiários, jornais impressos e diversos sites jornalísticos oferecem uma variedade de informações e posições políticas. Um noticiário deve ser imparcial transmitir um fato. Nos dias atuais, com um mundo cada vez mais polarizado, as pessoas tendem a adotar "um lado" com relação a qualquer assunto e este comportamento tem sido observado com relação à imprensa, seja pela opinião do jornalista, por orientação do editorial ou por interesse de terceiros.

1.3 OBJETIVOS

Uma notícia tendenciosa é, às vezes, bem suave para o interlocutor, e alcança seu objetivo de influenciar a opinião do leitor no mesmo sentido. Nos dias atuais, devido a quantidade de informações existentes, muitas pessoas sentem dificuldades em avaliar a ideia principal do conteúdo informado ou se existe alguma tendência, no caso deste trabalho, política. Este trabalho se motiva a estudar, através do processo de descoberta do conhecimento em textos e utilizando como classificador a Rede Neural Sem Peso WiSARD, uma arquitetura para automatizar o processo de identificação de origem e de polaridade em notícias políticas do Brasil.

As principais perguntas que este trabalho pretende responder são:

- É viável classificar automaticamente as fontes e a polaridade de artigos sobre política?

- Qual impacto da redução de dimensionalidade no desempenho do classificador?
- Quais são as vantagens e desvantagens dos classificadores WiSARD, SVM, *Naive Bayes*, *Gradient Tree Boosting* e Regressão Logística, em relação à precisão e eficiência/desempenho?
- Como avaliar a evolução temporal da polaridade em fontes de notícias?

O projeto aqui descrito tem como objetivos gerais:

- Propor uma arquitetura para a classificação fontes de notícias e identificação de polaridade em textos escritos em português;
- Avaliar o desempenho do modelo, ao utilizar a Decomposição em Valores Singulares;
- Utilizar a WiSARD, criada originalmente para o reconhecimento de imagens, como classificador de textos e avaliar seu desempenho; e
- Identificar fatos que expliquem a polaridade em um noticiário, através da avaliação de sua evolução temporal. Para isso, utilizar a ClusWiSARD, devido à sua propriedade de identificar padrões muito diferentes pertencentes à uma mesma classe.

1.4 CONTRIBUIÇÕES

As principais contribuições alcançadas com este estudo são as seguintes:

- **Viabilidade da WiSARD:** Identificação da viabilidade do uso da rede WiSARD para análise de dados textuais, por ser um modelo simples e com baixos

requisitos computacionais para seu funcionamento. Foi alcançada uma acurácia superior a 90%, desempenho comparável à Regressão Logísitca, que obteve melhor acurácia no comparativo entre os os classificadores WiSARD, SVM, *Naive Bayes*, *Gradient Tree Boosting* e Regressão Logística;

- **Avaliação dos ganhos devido à redução de dimensionalidade:** Avaliação prática do processo de Análise Semântica Latente para estimativa de tendência em textos políticos. Dentre as vantagens do uso de Análise Semântica Latente, destacamos a redução do tempo de treinamento dos classificadores e uma alta acurácia quando reduzida a dimensionalidade.

- **Estimação da evolução temporal de tendência:** Uma nova metodologia para avaliação de evolução temporal de polaridade em notícias sem uma classificação prévia.

1.5 TRABALHOS RELACIONADOS E ESTADO DA ARTE

1.5.1 Análise de Sentimentos em Notícias

Com relação às notícias curtas em português do Brasil, o trabalho de [35] aborda a mineração de textos através do processo de Análise Semântica Latente para realizar a identificação de emoções básicas (alegria, raiva, tristeza, desgosto, medo e surpresa) em notícias curtas. Neste trabalho, os grupos de palavras associadas a cada emoção foram dispostos na mesma dimensão das notícias do conjunto de treinamento. Para avaliação da emoção da notícia, foi utilizada a similaridade por cosseno, onde foi calculado o cosseno entre o vetor da notícia avaliada e cada vetor do conjunto de treinamento, obtendo-se como resposta a emoção do vetor no qual o cosseno foi maior.

O trabalho de [27] aborda a análise de sentimentos de textos curtos em tempo real. Neste trabalho é proposta uma forma de representar os termos que normalmente são encontrados nos picos do fluxo de dados, criando uma alternativa à ponderação TF-IDF. O trabalho se inspira na psicologia social e analisa as postagens ao vivo no debate de dois esportes populares no *Twitter*, futebol e futebol americano, gerando rótulos de forma automática.

A identificação de emoções em textos também á abordada no trabalho de [20], onde o corpus é formado por notícias que se enquadram em variadas categorias como mundial, nacional, política, policial e econômica. As notícias foram rotuladas manualmente de acordo com as emoções alegria, raiva, tristeza, desgosto, medo e surpresa. Após o pré-processamento, a redução de dimensionalidade é realizada com base no cálculo do ganho de informação pela entropia para todos os termos e então filtrados os mais relevantes. Para a classificação da emoções, é utilizado uma máquina de vetores de suporte

1.5.2 Análise de Polaridade em Notícias

O trabalho de [24] busca avaliar a polaridade (positiva, negativa ou neutra) de títulos de notícias de economia, disponíveis em endereços de *RSS Feeds*. No trabalho, é utilizado um software comercial para realizar o processo de descoberta de conhecimento em textos em português. Ainda são comparados os modelos estatístico, o baseado em regras para os termos e é proposto um modelo que possui técnicas estatísticas e regras para os termos.

Em [4] é proposto um modelo de identificação de polarização para redes sociais e é demonstrado que a presença de grupos polarizados pode ser detectada considerando a dependência entre as observações postadas e compartilhadas. Usando

uma matriz de usuários e *posts* como parâmetro, é proposta uma abordagem de fatoração de matrizes para descobrir a polarização. No trabalho, não é considerada a análise de sentimento, visto que uma declaração negativa sobre um determinado assunto não significa que seja oposta. São explorados diferentes graus de polarização e comparada a qualidade de separação (de *tweets* de polaridade oposta) através de diferentes algoritmos usando dados reais coletados do *Twitter*.

O trabalho de [7] aborda a identificação de polaridade em notícias políticas americanas através da rede social *Facebook* [5]. Sua avaliação consistem em explorar a correlação entre os usuários que declaram sua ideologia política (liberal, conservador ou moderado) em seu perfil na rede social e as notícias que são compartilhadas ou clicadas por eles. Para identificação da categoria de notícias, é utilizado o classificador SVM. O trabalho aborda o delineamento do perfil ideológico dos principais sites de notícias americanos. Também abordada a relação entre os usuários de pensamentos ideológicos distintos.

1.5.3 Análise Conjunta de Polaridade e Sentimento

O trabalho de [8] aborda a identificação de polaridade analisando *posts* de *blogs* políticos e suas seções de comentários de diferentes comunidades. O trabalho apresenta uma forma para determinar a polaridade do sentimento dos comentários nos *blogs* e a tarefa de prever a polaridade com base no conteúdo da postagem do *blog*. Ainda é apresentado um modelo, o MCR-LDA, que identifica os tópicos e as respostas que eles apresentam em diferentes *blogs*, mesmo sendo de polaridades opostas.

[5] Site do Facebook: `http://www.facebook.com`

1.5.4 Depuração de Mecanismos de Classificação

Um grande desafio no domínio de classificação de texto (e de aprendizado por máquina em geral) consiste em depurar o resultado final obtido. Quais os fatores que levaram a uma determinada classificação? Por que o classificador gerou uma determinada resposta? Qual teria sido a resposta caso os dados fossem distintos? A resposta a estas perguntas é fundamental para aumentar a confiança no uso de métodos de aprendizado por máquina [49].

O uso da WiSARD traz algumas vantagens no sentido de facilitar a depuração dos resultados obtidos.

1. **simplicidade:** o fato de a WiSARD ser muito simples facilita o entendimento de quais fatores impactam seus resultados. Em particular, o estado da WiSARD consiste de vetores binários, que embora não sejam imediatamente interpretáveis, mapeiam diretamente as entradas nas saídas.

2. **uma entrada pode estar associada a vários rótulos:** cada discriminador distingue cada amostra como condizente ou não a uma determinada classe. Assim sendo, pode-se associar múltiplos rótulos a uma mesma amostra. Esta propriedade foi explorada, por exemplo, em [45]. Neste trabalho, focamos apenas na classe mais representativa, mas vislumbramos como trabalho futuro fazer a distinção entre amostras que possam ser classificadas com múltiplos rótulos, identificando suas propriedades (e.g., amostras neutras versus polarizadas).

3. **identificação dos papeis dos atributos:** aliando-se os dois pontos acima (simplicidade e possibilidade de associar múltiplos rótulos a uma mesma entrada), pode-se gradativamente alimentar diferentes atributos à WiSARD, e

verificar como a resposta dos discriminadores varia na medida em que os atributos são adicionados. Assim, pode-se identificar o papel de cada atributo, e sua importância, gerando uma explicação para a resposta final obtida com o método. Essa estratégia foi utilizada em [45] e pretendemos, também em trabalhos futuros, adotar algo similar no sentido de alimentar diferentes palavras do *bag-of-words*, de forma incremental, à WiSARD, e identificar como que a discriminação entre partidos ou tendências políticas varia em função das novas palavras apresentadas. Note também que em [45] não foi adotada a técnica de *bleaching*, o que por si só pode ser uma extensão interessante, conforme considerado no presente trabalho.

1.5.5 Resumo dos Trabalhos Relacionados

Neste trabalho consideramos o uso da arquitetura WiSARD para classificação de notícias políticas. Não é de nosso conhecimento nenhum trabalho anterior que tenha utilizado tal arquitetura para fins de detecção de polaridade em notícias políticas. Também não é de nosso conhecimento nenhum outro trabalho que tenha avaliado aspectos temporais da evolução da polaridade de um meio de comunicação.

1.6 ORGANIZAÇÃO DA DISSERTAÇÃO

No Capítulo 2, são apresentados os conceitos básicos do processo de mineração de textos, onde apresentamos de forma resumida as etapas que compõe o processo juntamente com uma abordagem das Redes Neurais Sem Peso, da WiSARD e da ClusWiSARD. O Capítulo 3 apresenta todo o processo de descoberta de conhecimento em textos na forma que foi conduzido neste trabalho. Com a arquitetura definida, o Capítulo 4 apresenta os diversos experimentos realizados e seus

respectivos resultados e no Capítulo 5, finalizamos o trabalho com as conclusões da pesquisa e apresentamos possíveis trabalhos futuros.

Com a finalidade de apoiar aos conceitos básicos, no Apêndice A falamos sobre os termos abordados e outras considerações, no Apêndice B apresentamos um detalhamento sobre o funcionamento da WiSARD e a ClusWiSARD. Por fim, no Apêndice C abordamos os classificadores utilizados para comparação com a WiSARD e as métricas utilizadas para avaliação.

2 CONCEITOS BÁSICOS

Neste capítulo apresentamos um breve resumo das etapas do processo de mineração de textos. Apresentamos também uma introdução à Rede Neural Sem Peso WiSARD. No Apêndice B apresentamos mais detalhes sobre a WiSARD e a ClusWiSARD. No Apêndice C temos uma descrição dos demais classificadores considerados e das métricas para comparação.

2.1 COLETA DE DADOS

Consideramos o ponto de partida da mineração de textos, onde se identificam possíveis fontes de documentos relevantes, que podem ser em bancos de dados, arquivos locais ou na *web*. Quando os dados estão na *web*, a coleta pode ser efetivada com um algoritmo que realiza varredura em uma página específica ou em várias, de forma recorrente ou não, armazenando os dados desejados. Este algoritmo é denominado *crawler*. Uma vez encontrada a fonte, devem ser coletados os dados relevantes e armazenados na coleção de documentos que será analisada.

2.2 PRÉ-PROCESSAMENTO

Após a coleta e a definição do modelo de representação de textos, os dados brutos passam por uma série de tratamentos para que possam ser utilizados pelos algoritmos de mineração de textos. Esta fase é a mais onerosa do processo de mineração de textos, pois são realizados vários experimentos como podemos citar:

- **remoção de acentos**, que reduz a quantidade de caracteres;
- ***tokenização***, que identifica os termos dentro da sequência de caracteres;
- **remoção de *stopwords***, que reduz a quantidade de termos eliminando palavras que não contribuem para descrever o documento; e
- ***stemming***, que reduz os termos ao seu radical, considerando variações de um termo uma única vez.

Ainda existem diversas técnicas que são utilizadas no pré-processamento para alcançar a representação adequada.

2.3 TRANSFORMAÇÃO DOS DADOS

Nesta fase, levamos em consideração que nem todas as palavras presentes em um documento possuem a mesma importância. Com exceção das *stopwords*, os termos mais frequentemente utilizados costumam ter significado mais importante. Após a etapa do pré-processamento, para cada termo pertencente ao índice, é atribuído um peso, que pode ser definido de diversas formas. As mais comuns são baseadas em cálculos simples de frequência do termo.

2.4 MINERAÇÃO

Na mineração, ocorre a busca efetiva sobre o conhecimento acerca dos dados. Podem ser utilizadas técnicas de redução de dimensionalidade dos dados. Uma técnica que pode ser adotada é denominada Análise Semântica Latente, que é um processo para identificar correlações latentes entre os termos e os documentos. Ela

não é considerada como técnica de processamento de linguagem natural, por utilizar somente os termos, sem bases criadas por humanos ou regras [32]. A partir da matriz de co-ocorrência dos termos e dos documentos, busca-se identificar novas relações [18]. É um campo oriundo de Recuperação da Informação e tem ampla utilização categorização de documentos por similaridade [11].

Ainda são aplicados algoritmos de aprendizado de máquina. É importante ressaltar a dificuldade em se encontrar um algoritmo adequado, pois devido a grande diversidade existente muitas vezes é necessário realizar diversos testes. Neste trabalho abordamos a Rede Neural sem Peso WiSARD, a ClusWiSARD e outros classificadores.

2.5 REDES NEURAIS SEM PESO

As Redes Neurais Sem Peso (RNSP) começaram a ser estudadas tendo como principal motivação as dificuldades enfrentadas pela comunidade para implementação em hardware dos pesos das conexões de neurônios do tipo McCulloch-Pitts [38]. O primeiro trabalho voltado para as RNSPs, foi o método n-tuplas de [10], um estudo voltado para o reconhecimento de caracteres alfanuméricos. O projeto foi composto de um mosaico de fotocélulas de tamanho 10 x 15, com agrupamentos em n-tuplas aleatórias, onde n=2, formando 75 pares exclusivos conforme pode ser observado na Figura 2.1.

Cada tupla pode poderia receber 4 valores distintos (00, 01, 10 e 11). No mosaico eram projetados os caracteres, de modo que um valor binário era atribuído a cada fotocélula de acordo com a iluminação da mesma. Desta forma, cada par previamente selecionado atribuía o valor à respectiva tupla, conforme pode ser observado na Figura 2.2.

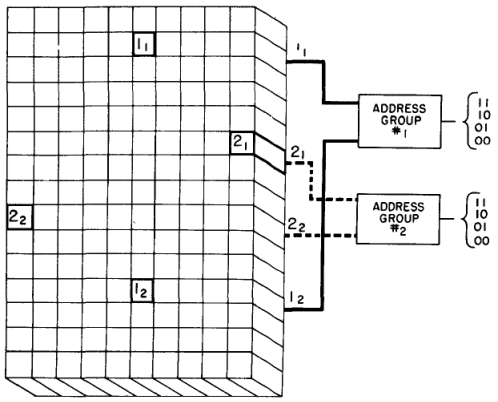

Figura 2.1: O fotomosaico e dois dos pares de fotocélulas escolhidas aleatoriamente. Os quatro grupos digitais à direita são os quatro estágios possíveis de cada par de fotocélulas [10].

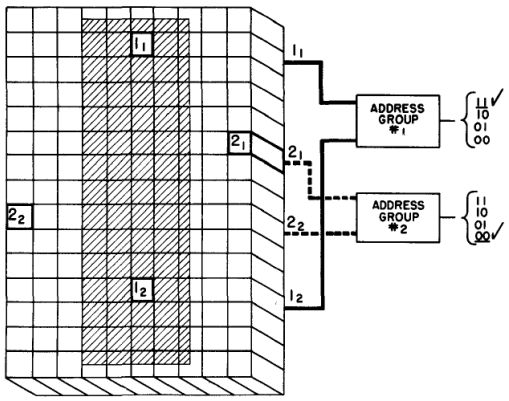

Figura 2.2: O sistema que aprende a letra I em uma posição central. Apenas dois dos 75 pares são mostrados [10].

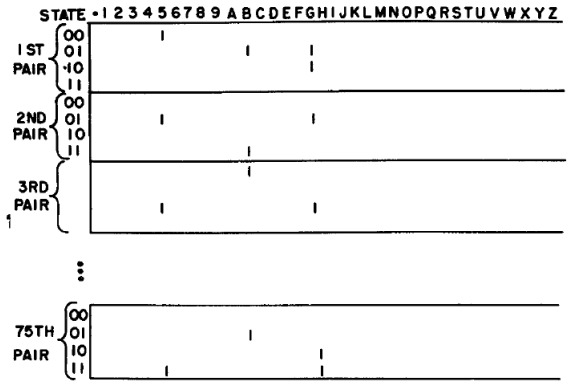

Figura 2.3: Matriz de memória com os caracteres 5, B e G [10].

Além do mosaico, foi utilizada uma matriz de memória onde a quantidade de linhas era a quantidade de combinações das n-tuplas, neste caso 300 (4 combinações x 75 n-tuplas) e a quantidade de colunas eram a representação dos caracteres alfanuméricos, num total de 36 (um ponto, 9 números e 26 letras). Desta forma, quando um caractere era apresentado ao mosaico, o valor lógico '1' era atribuído na matriz de memória conforme pode ser observado na Figura 2.3.

Para o reconhecimento, assim que apresentado o padrão desconhecido, foi calculada a soma dos valores referente a cada caractere previamente apresentado no treinamento e, após a comparação, o caractere referente à maior pontuação era atribuído ao padrão apresentado, conforme pode ser observado na Figura 2.4. É importante ressaltar que quando existia o empate entre as pontuações máximas, o vencedor era escolhido aleatoriamente.

As Redes Neurais Sem Peso (RNSP) e as Redes Neurais com Peso (RNA) são inspiradas no sistema nervoso humano, porém, com abordagens distintas. Podemos citar algumas diferenças entre uma RNA e uma RNSP. Com relação ao armazenamento da informação, as RNAs detém a informação nos pesos das ligações e as

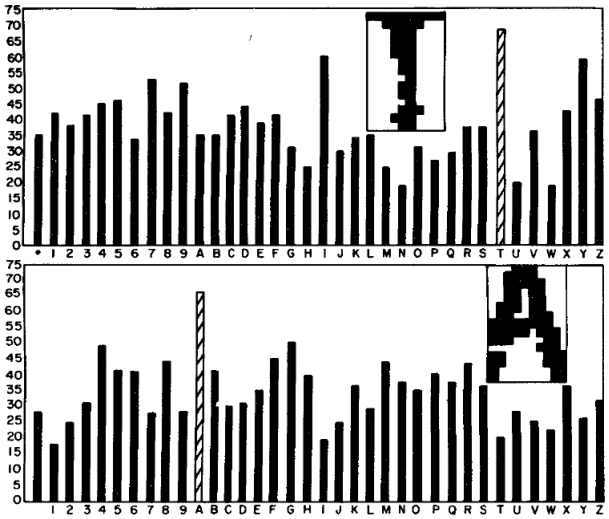

Figura 2.4: Comparativo para o reconhecimento [10].

RNSPs armazenam a informação na memória RAM, também denominada neurônio RAM (tabela endereçável) [42]. Podemos observar outras diferenças abaixo:

1. A RNAs utilizam valores reais e as RNSP utilizam valores discretos;
2. Os neurônios das RNAs computam funções linearmente separáveis e as RNSPs computam funções booleanas; e
3. Individualmente, os neurônios das RNAs podem generalizar e os neurônios RAM podem generalizar somente em nível de rede, em outra palavras, quando estão agrupados.

Uma estrutura com K neurônios RAM conectados com um padrão de entrada é denominada discriminador e tem a finalidade de reconhecer uma única classe. No discriminador, cada RAM aprende parte do padrão de entrada. Nessa estrutura ainda tem um dispositivo somador, que realiza a contagem das respostas do neurônios RAM e retorna o total de neurônios que reconheceram o padrão apresentado.

2.6 REDE NEURAL SEM PESO WISARD E A CLUSWISARD

A WiSARD é uma rede neural sem peso composta por um grupo de discriminadores. Cada discriminador é responsável pelo reconhecimento de uma única classe. Ela utiliza vetores binários como entrada e o armazenamento da informação é realizado pelos neurônios RAM, que não são capazes de generalizar de forma isolada, porém, conseguem generalizar quando compõe o discriminador [42].

A ClusWiSARD é uma variação da WiSARD, que busca melhorar o desempenho da WiSARD ao utilizar discriminadores como *clusters*, com a finalidade prevenir a classificação incorreta da WiSARD por conta do treinamento de padrões muito distintos de uma mesma classe. A ClusWiSARD utiliza um grupo de discriminadores por classe, fazendo com que os padrões apresentados no treinamento sejam absorvidos por discriminadores que melhor os representem [14].

2.7 ANÁLISE

Para analisar o resultado do processo de mineração de textos, são utilizadas diversas métricas de avaliação de desempenho. Algumas que podemos citar:

- **Acurácia**, percentual de quantos documentos foram classificados corretamente pelo classificador dentre o total de documentos avaliados;
- **Precisão**, representa o percentual de documentos classificados corretamente como positivo sobre o total de documentos classificados como positivo; e
- **Revocação**, representa o percentual de documentos classificados como positivo dentre todos os documentos que realmente são positivos.

3 METODOLOGIA

A metodologia proposta neste trabalho tem por objetivo definir as etapas para o processo de mineração de textos, com a finalidade de realizar a classificação de textos dentre uma coleção.

Com a finalidade de se obter um maior conhecimento e uma maior destreza prática sobre a metodologia de mineração de textos, foram implementadas com auxílio de bibliotecas as etapas de coleta de dados, pré-processamento, cálculo de relevância, redução de dimensionalidade e análise dos resultados. Todas as etapas do processo de mineração de textos foram implementadas na linguagem de programação Python na versão 2.7, 64 bits.

A metodologia adotada é composta por 5 etapas: a coleta de documentos, o pré-processamento, a transformação dos dados, a mineração e a fase de análise. É importante ressaltar que é necessário realizar esta sequência de etapas durante o processo de mineração de textos. Caso o resultado final não seja satisfatório, alguma etapa pode ser modificada. Após a modificação, é necessário que as etapas subsequentes sejam realizadas.

3.1 COLETA DE DOCUMENTOS

Como o estudo de caso deste trabalho é referente a classificação textual, o passo inicial foi a aquisição uma base textual para a extração de conhecimento. Devido à dificuldade de se encontrar uma base de dados pré-classificada que contivesse notícias políticas em português do Brasil, foi necessário buscar na *web* os textos para

se montar o a base de dados do trabalho e classificá-los de forma manual.

O primeiro passo para a coleta, foi a definição do tipo de coletor a ser utilizado. Como a necessidade era a de se obter notícias políticas, foi definida a implementação de um *crawler* para obter textos das páginas dos respectivos partidos políticos. Foram escolhidos quatro partidos políticos para a coleta de notícias em seus portais: o Partido do Movimento Democrático Brasileiro (PMDB), Partido da Social Democracia Brasileira (PSDB), Partido dos Trabalhadores (PT) e o Partido Socialismo e Liberdade (PSOL).

Após a adequação do *crawler* à estrutura de cada página, as notícias foram coletadas e armazenadas individualmente em arquivo no formato TXT, onde o nome do arquivo recebeu a manchete da notícia e o corpo da notícia ficou armazenado no arquivo, e separadas por pastas referentes a cada partido. Foram coletadas 1147 notícias, dos anos de 2016 e 2017 e sem qualquer seleção de assuntos ou temas.

3.2 PRÉ-PROCESSAMENTO

Esta fase tem o objetivo de preparar os dados, reduzir e transformar os textos 'brutos'. Dentre todas as fase do processo de mineração, esta é a que consome maior parcela de tempo. Neste trabalho será abordada a análise estatística dos textos, onde para cada termo é atribuído um peso. Para a representação dos documentos, utilizamos o modelo *bag-of-words* para a representação dos documentos. Cada documento é representado por um vetor onde a quantidade de termos é determinada pela quantidade de termos distintos no conjunto de documentos e cada posição é relacionada ao peso de um termo dentro do documento. Apesar da escolha sobre a abordagem estatística dos textos, a fase de pré-processamento, é essencial para 'limpar' os dados.

Ne metodologia utilizada, foi elaborada uma função somente para o pré processamento, que possui como parâmetro um arquivo de extensão TXT por vez a ser processado. Cada notícia é tratada inicialmente como uma sequência de caracteres, sem qualquer correlação lógica ou semântica. Inicialmente, são removidos os caracteres especiais, dígitos, acentos e pontuações, pois não serão úteis ao processamento. Em seguida, todos os caracteres são convertidos para caixa baixa.

Após a remoção dos caracteres e integração de caixa, a sequência de caracteres é convertida para *tokens*, ou seja, são identificadas as palavras em si. Com as palavras identificadas, é utilizado um dicionário para a remoção das *stopwords*, que no português do Brasil são compostas por artigos, pronomes, preposições e verbos auxiliares, fazendo com que se reduza consideravelmente a quantidade de palavras do texto. Por abordarmos o modelo de representação *bag-of-words*, consideramos a palavra 'não' como uma *stopword*. Devemos estar atentos para que informações úteis não sejam descartadas nesta fase.

Logo em seguida, é realizado o *stemming*, onde as palavras do texto são reduzida ao seu radical. Em outras palavras, consolidamos a relação de várias palavras para uma único termo que pode substituí-las sem modificar o contexto. Esta etapa proporciona uma redução considerável na quantidade termos candidatos a índice dos documentos. Esta etapa é muito importante para as fases posteriores.

Como o objetivo de aproveitar termos compostos, foi utilizada a técnica de bigramas. Desta forma, um termo composto importante pode se destacar no cálculo de relevância dos termos. A execução das etapas citadas anteriormente resultou em uma representação de termos correlacionados a cada documento. O conjunto de termos ainda persistentes ao pré-processamento é denominado por índice de cada documento. Após o pré processamento, obtivemos vetores com 11485 posições.

3.3 TRANSFORMAÇÃO DOS DADOS

Este trabalho aborda uma análise estatística dos dados, através da representação pelo modelo *bag-of-words*. Com a definição dos índices dos documentos, torna-se necessária a transformação do conjunto de termos e documentos para a representação espaço-vetorial. Nesta abordagem, cada componente do vetor-documento traduz numericamente a importância semântica de um termo presente no mesmo.

Considerando n documentos indexados e m termos podemos representar a matriz termo-documento A de ordem $m \times n$. Os vetores-documento estão dispostos como colunas na matriz A e cada elemento $a_{i,j}$ representa a frequência ponderada que o termo i ocorre no documento j.

Cada termo em um documento possui um determinado grau de relevância ao contexto do documento em geral. Utilizando a representação *bag-of-words*, foi escolhido o modelo TF-IDF para representar a relevância de cada termo, visto que a ponderação dos termos apresenta uma melhor performance na mineração de textos em comparação ao modelo booleano, ou seja, contando somente a existência do termo no documento ou não. Após o cálculo da frequência de cada termo (TF) e da frequência inversa de cada documento (IDF), foi criada uma matriz termo-documento contendo em suas linhas, os termos contidos em todos os documentos e em suas colunas, a representação de cada documento. Caso um termo não esteja contido em um determinado documento, sua posição na matriz terá o valor 0 e no caso contrário, terá o valor resultante do cálculo de relevância.

3.4 MINERAÇÃO E REPRESENTAÇÃO EM VALORES BINÁRIOS

Nesta fase acontece o refinamento dos resultados das fases anteriores. Dentre as abordagens de seleção de características e extração de características para redução de dimensionalidade dos dados, foi escolhida a de extração de características. Para a execução deste fase foi escolhida a Decomposição em Valores Singulares (SVD).

A matriz termo-documento ponderada contém toda informação estrutural entre os termos e os documentos do conjunto, porém, de forma latente. Em outras palavras, termos distintos podem se relacionar com outros termos e com documentos onde eles não são mencionados. A partir da matriz termo-documento ponderada, podemos extrair relações geométricas e algébricas entre os termos e documentos (vetores) para avaliar semelhanças e diferenças semânticas de conteúdo através da SVD.

A Decomposição em Valores Singulares será utilizada para localizar informação latente na matriz termo-documento ($termos \times documentos$), baseada na co-ocorrência de palavras nos documentos. Baseado na escolha dos maiores valores singulares, podemos reduzir a dimensão da matriz, como forma de aproximar a matriz A por combinações lineares e obtendo as melhores relações entre os termos e os documentos, além de conseguir reduzir o 'ruído' da matriz original, isto é descartar informações insignificantes, que na prática, são associados aos menores valores singulares resultantes da decomposição.

Seja ki o posto da matriz termo-documento. A SVD pode ser interpretada como o mapeamento do espaço de A em um espaço conceito (reduzido) de k dimensões, as quais são linearmente independentes. Neste novo espaço, os vetores de termos em U tem k entradas, cada um dando a ocorrência do termo i em um dos k conceitos. Da mesma forma, os vetores de documentos em V revelam a relação en-

tre o documento j com cada conceito k. Usualmente formaliza-se o espaço conceito como

$$A_k = U_k \Sigma_k V_k^T$$

onde A_k é a representação da matriz termo-documento reduzida, U_k representa a relação entre os termos e os conceitos, V_k representa a relação entre o documentos e os conceitos e Σ_k representa a intensidade de cada conceito.

Para a escolha do valor de k que melhor representa a aproximação de A, foi utilizada uma forma empírica para encontrar o valor. Foram plotados os valores Σ e foi observado o ponto onde os valores de Σ_k se estabilizam, ou seja, o 'cotovelo' do gráfico. Foi observado que os valores de k posteriores a este ponto acrescentavam uma quantidade de informação insignificante para extração de informação. Na Figura 3.1, podemos observar e concluir sobre os valores de k que melhor representam a aproximação em relação à matriz original, no caso, o valor de $k = 20$ satisfaz o valor que estamos buscando. No experimento, reduzimos a representação de 11485 termos para algumas dimensões, com variações de tamanho entre 20 e 100 a fim de avaliar o impacto nos resultados do classificador.

Ressaltamos uma grande vantagem na execução deste procedimento, visto que possuíamos originalmente 11485 termos e após a redução dimensional, cada documento pôde ser representado por 20 conceitos, mantendo-se uma acurácia significativa, fato que afeta diretamente na redução do custo computacional.

Uma vez que a matriz A incorpora (por construção) todas as associações estruturais entre os termos e os documentos, observa-se que, para uma determinada coleção de treinamento, o produto AA^T caracteriza todas as coocorrências entre as palavras e A^TA caracteriza todas as coocorrências entre os documentos.

Para a tarefa de classificação dos documentos, foi realizada a expansão de

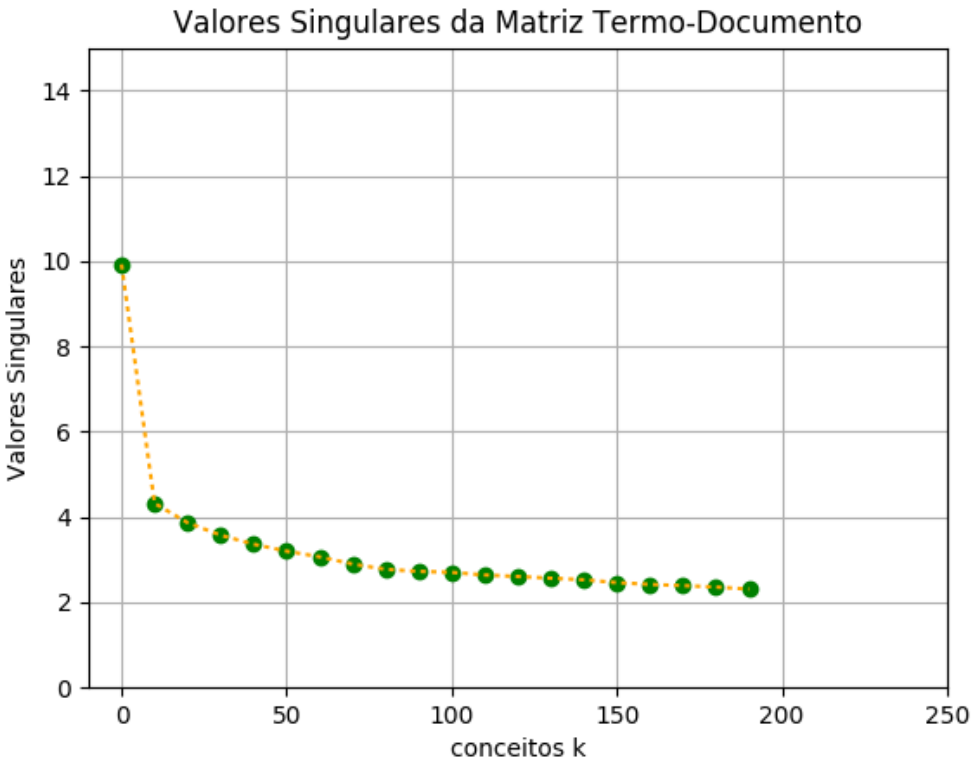

Figura 3.1: Valores de Σ.

AA^T, alcançando

$$AA^T = U\Sigma^2 U^T$$

o que significa que a célula (i, j) de AA^T pode ser obtida através do produto entre $u_i\Sigma$ e $u_j\Sigma$. A matriz unitária U_k foi utilizada para realizar mapeamento dos vetores referentes aos documentos (d) para os conceitos escolhidos. Cada exemplo do conjunto de treinamento e de testes foi representado por

$$\hat{d} = d^T U_k$$

onde d^T representa o vetor do documento e $\hat{d}$ é a nova representação do documento. Com a nova representação, foi necessário o ajuste dos vetores para o formato de entrada da WiSARD.

Conforme já foi abordado anteriormente, o WiSARD somente recebe como entrada vetores binários. A matriz resultante do processo de redução de dimensionalidade possui vetores contínuos. Esses valores inviabilizam a utilização do WiSARD. A partir desta constatação, torna-se necessária a conversão dos vetores contínuos para vetores de bits.

Para a conversão para o sistema numérico binário foi utilizada uma forma de representação similar a um 'termômetro'. Nesta forma de representação, foi definido um valor de V_{bin} bits para representar cada valor da matriz de números contínuos. Foi observado que todos o valores absolutos dos números a serem convertidos variavam entre 0 e 1, com muitas casas decimais. Para a definição do tamanho de V_{bin}, foi utilizada a soma dos valores absolutos do maior e do menor elemento da matriz, multiplicada por um número de escala, que no caso foi 100. Seja $H(V_{bin})$ o tamanho de V_{bin}, em bits. Então,

$$H(V_{bin}) = (|\min(A_k)| + |\max(A_k)|)e \tag{3.1}$$

onde H é o tamanho de V_{bin}, $|\min(A_k)|$ é o menor valor absoluto da matriz, $|\max(A_k)|$ é o maior valor absoluto da matriz e e é uma escala escolhida empiricamente. Um

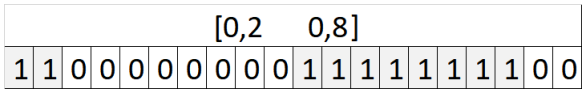

Figura 3.2: Representação para o vetor $[0,2 \quad 0,8]$

exemplo pode ser observado na Figura 3.2, onde é criada uma nova representação para um vetor de duas posições, com e=10.

Cada valor contínuo em $\hat{d}$ é somado com o valor de $|\min(A_k)|$ e então multiplicado pelo número de escala (100). Este valor encontrado é a quantidade de números '1' que V_{bin} conterá. A partir de então, cada vetor binário será concatenado, formando a representação de um documento. Na Figura 3.2 pode ser observado um exemplo de representação binária no modelo 'termômetro' de um vetor de números contínuos, com escala e tamanho 10.

Uma vez que todos os exemplos são representados desta forma, os vetores de documentos podem treinar o WiSARD e a rede neural sem peso decidirá a quais categorias podem ser atribuídas as amostras de teste.

3.5 ANÁLISE

A última etapa do processo de mineração de textos e responsável pela avaliação e interpretação dos padrões extraídos. Esta etapa visa constatar se o objetivo almejado foi alcançado, ou se todas ou algumas etapas no processo necessitam ser refeitas. Os padrões descobertos podem ser avaliados pelo usuário final, especialista do domínio e analista de dados, com o intuito de validar o conhecimento obtido [21].

Foi utilizada a biblioteca PyWANN [47], que é uma implementação da RNSP

WiSARD escrita para a linguagem de programação Python. Para utilização da WiSARD como classificador textual, foi necessária a utilização vetores rotulados para realizar a etapa de treinamento. Utilizaram-se os vetores binários, juntamente com um dos vetores de rótulos, obtidos a partir os arquivos TXT armazenados na etapa de coleta de dados, onde cada rótulo recebeu a nomenclatura da sigla do partido político referente ao documento (PMDB, PSDB, PT e PSOL). Para identificação de polaridade, foi utilizado um outro vetor de rótulos com agrupamento relativo à cada partido, conforme o comentado na Seção 1.1.

Foi utilizada a rede neural com 8, 16, 32 e 64 bits de endereçamento de memória. Para a avaliação dos resultados obtidos com o WiSARD, foi utilizado o método *K-fold* com 90% da massa de dados como treinamento e 10% para testes.

4 EXPERIMENTOS E RESULTADOS

4.1 OBJETIVO DOS EXPERIMENTOS

Com o objetivo de avaliar a performance da rede WiSARD para a classificação de documentos e identificação de polaridade em notícias, foram realizados alguns experimentos.

Os experimentos envolveram a classificação e a identificação de polaridade em notícias. Foram avaliadas duas formas de abordagem sobre a mineração de textos, na primeira forma, foi avaliada a notícia completa, com todos os termos utilizados para descrever a ideia da informação. No segundo formato de avaliação, foi observado o processo de mineração utilizando como base de dados somente as manchetes das notícias, devido à utilização de termos que buscam descrever todo o conteúdo de forma sucinta e enfática e também em virtude de muitas pessoas compartilharem reportagens em redes sociais sem ler seu conteúdo [25].

Todos os experimentos foram realizados utilizando-se a linguagem de programação Python, na versão 2.7 - 64 bits, rodando na plataforma Windows 10 Home. Como apoio à implementação, as principais bibliotecas para Python foram utilizadas: Scikit-Learn [43], PyWANN [47], Numpy [57] e Matplotlib [30]. A especificação principal do hardware utilizado foi um *notebook*, com processador Intel Core i7-6500U e 8Gb de memória RAM.

Tabela 4.1: Distribuição de notícias por partido

PARTIDO	QUANTIDADE DE NOTÍCIAS
PMDB	292
PSDB	315
PT	388
PSOL	152

Tabela 4.2: Distribuição de notícias por polaridade

Polaridade	Quantidade de Notícias
PMDB/PSDB	607
PT/PSOL	540

4.2 DADOS UTILIZADOS

Esta seção detalha as principais características da base de dados utilizada para avaliação da rede WiSARD. Foram coletadas notícias do ano de 2016 e 2017. As notícias dos sites dos seguintes partidos foram utilizadas como base de dados textuais: Partido do Movimento Democrático Brasileiro (PMDB - http://www.pmdb.org.br), Partido da Social Democracia Brasileira (PSDB - http://www.psdb.org.br), Partido dos Trabalhadores (PT - http://www.pt.org.br) e o Partido Socialismo e Liberdade (PSOL - http://www.psol50.org.br).

Quatro *arrays* foram utilizados para a avaliação. No primeiro *array* cada posição possuía o texto completo da notícia. O segundo *array* possuía em cada campo a sigla do partido político referente à notícia. O terceiro *array*, continha a polaridade de cada partido (PMDB/PSDB ou PT/PSOL) e no quarto *array*, cada posição possuía a manchete da notícia. Na Tabela 4.1 podem ser observados os quantitativos de notícias coletadas separadas por partidos políticos e na Tabela 4.2 podem ser observadas o quantitativo das notícias separadas por polaridade.

4.3 COMPARAÇÃO ENTRE CLASSIFICADORES

Este experimento foi demonstrado em [16] e realiza uma avaliação através do processo de mineração de texto utilizando o corpo da notícia como base de dados. Foram utilizadas notícias relacionadas aos partidos PT e PMDB, do ano de 2016. Após a coleta e categorização dos documentos dos sites das partes, seu pré-processamento foi realizado. Foram realizadas variações nos agrupamentos dos *tokens*, utilizando termos isolados, bigramas e trigramas. Para a redução de dimensionalidade, estratégia adotada foi aplicar a técnica de atribuição de pesos TF-IDF aos termos. Um limite estabelecido empiricamente θ foi utilizado para selecionar somente os termos mais significativos. Para construir a entrada binária, se o peso do termo fosse maior ou igual a θ, o valor considerado seria 1 e 0, caso contrário. Foi realizada uma comparação entre o classificador WiSARD e os classificadores SVM, *Naive Bayes*, Regressão Logística e *Gradient Tree Boosting*.

A tabela 4.3 retrata uma comparação entre o desempenho de algumas variações das arquiteturas do WiSARD (quantidade de bits de endereçamento 4, 8, 16 e 32 bits), com e sem *bleaching*, e os classificadores citados anteriormente. O WiSARD mostrou resultados próximos ao da Regressão Logística, que obteve o melhor resultado e resultados superiores aos classificadores SVM, *Naive Bayes* e *Gradient Tree Boosting* tanto na precisão como no desvio padrão.

Tabela 4.3: Comparativo entre classificadores [16]

Modelo	Unigramas		Bigramas		Trigramas	
	Ac	DP	Ac	DP	Ac	DP
Regressão Logística	88%	4%	95%	2%	94%	3%
WiSARD 4 bits com *bleaching*	88%	5%	93%	3%	95%	2%
WiSARD 8 bits com *bleaching*	88%	5%	93%	3%	94%	3%
SVM (kernel linear)	87%	4%	93%	4%	92%	4%
WiSARD 16 bits com *bleaching*	89%	6%	93%	6%	93%	4%
WiSARD 16 bits	89%	4%	92%	5%	93%	3%
WiSARD 8 bits	90%	6%	91%	3%	94%	3%
WiSARD 4 bits	91%	4%	91%	4%	94%	3%
GB (estimators=150, learning rate=0.01, depth=5)	83%	8%	86%	4%	85%	5%
Bernoulli Naive Bayes	80%	5%	85%	5%	86%	5%
WiSARD 32 bits com *bleaching*	83%	6%	84%	4%	93%	3%
WiSARD 32 bits	82%	6%	82%	6%	93%	4%

4.4 CLASSIFICAÇÃO E IDENTIFICAÇÃO DE POLARIDADE

4.4.1 Classificação Baseada no Corpo da Notícia

O corpo da notícia é onde existe todo conteúdo da informação, através dos termos relevantes, podemos identificar a essência do mensagem. Por conter muitos termos, a tarefa de realizar a classificação se facilita. No corpo, há um detalhamento maior dos fatos, de modo a destacar os detalhes mais importantes, fundamentais à compreensão do interlocutor. O *dataset* com os textos das notícias, após a fase de pré-processamento, possuía 11485 termos distintos e 1147 documentos.

Após a transformação dos dados em sua forma ponderada, foram realizados testes com diversas dimensões reduzidas. Foi avaliado o comportamento do classificador 10, 20, 30, 40 e 50 dimensões. Após a binarização das dimensões reduzidas, foi atribuído um rótulo para cada notícia de acordo com sua fonte partidária. Para avaliação de polaridade, foram atribuídos os rótulos da acordo com o grupo de sua fonte partidária. Foram avaliados os tamanhos de endereçamento com 8, 16, 32 e 64 bits para memória RAM da rede WiSARD.

Observamos que nas Figuras 4.1 e 4.2, utilizando o processo de validação *k-fold*, a acurácia obtida com a WiSARD com 64 bits de endereçamento ficou próxima a 90% do modelo adotado decai ao se reduzir o tamanho dos bits de endereçamento de memória. Este fato explica o aumento do potencial de generalização do modelo relacionado ao tamanho da memória. Com relação à quantidade de dimensões, apesar de utilizarmos dimensões de tamanho entre 10 e 50, as dimensões que apresentaram desempenho superior foram as de tamanho 30 e 20, fato que corrobora o apresentado sobre a melhor dimensão a ser utilizada, na Seção 3.4.

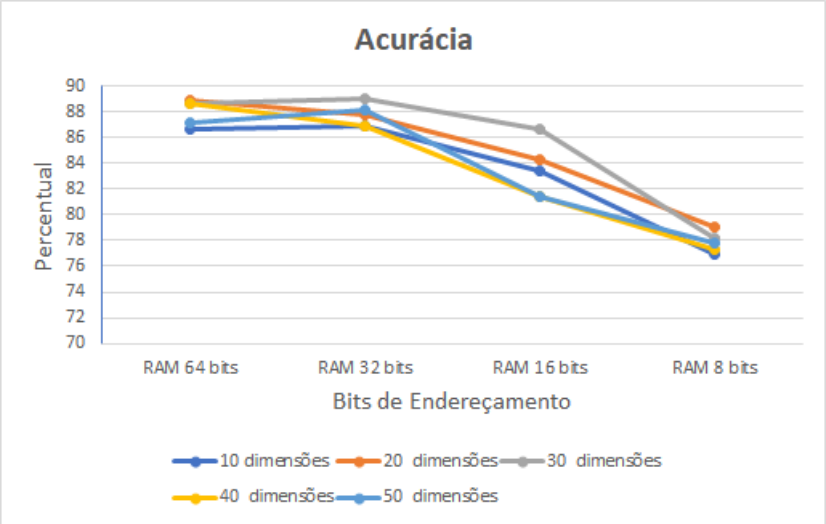

Figura 4.1: Acurácia relativa à classificação por polaridade baseada no corpo da notícia

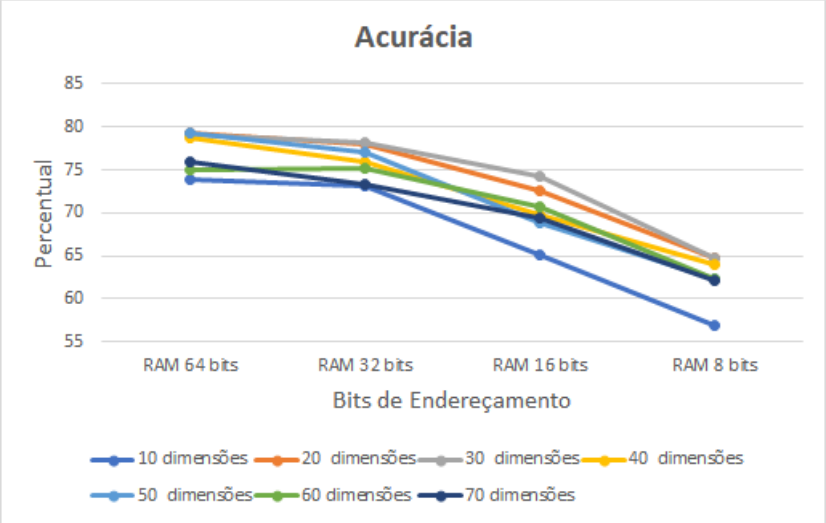

Figura 4.2: Acurácia relativa à classificação por partido baseada no corpo da notícia

Nas Figuras 4.3 e 4.4, observamos que o tempo de processamento cresce de

forma linear de acordo com a quantidade de dimensões dos dados utilizados no WiSARD.

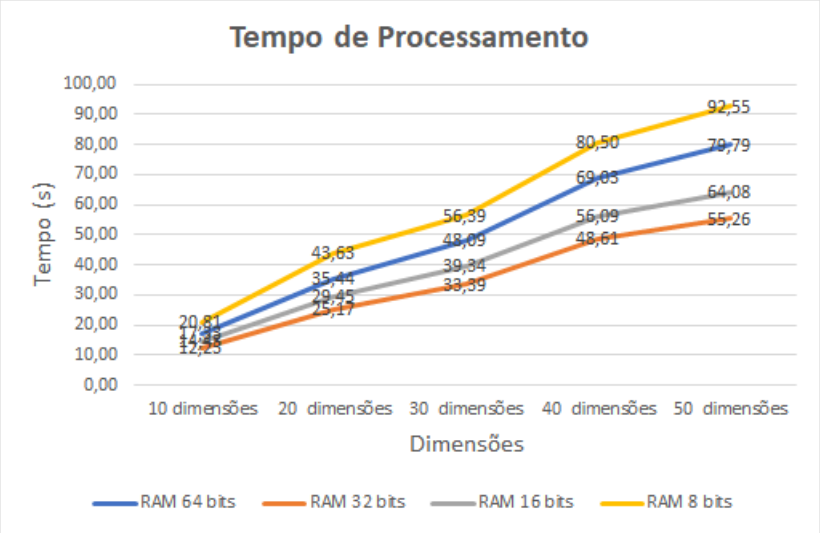

Figura 4.3: Tempo de processamento relativo à classificação por polaridade baseado no corpo da notícia

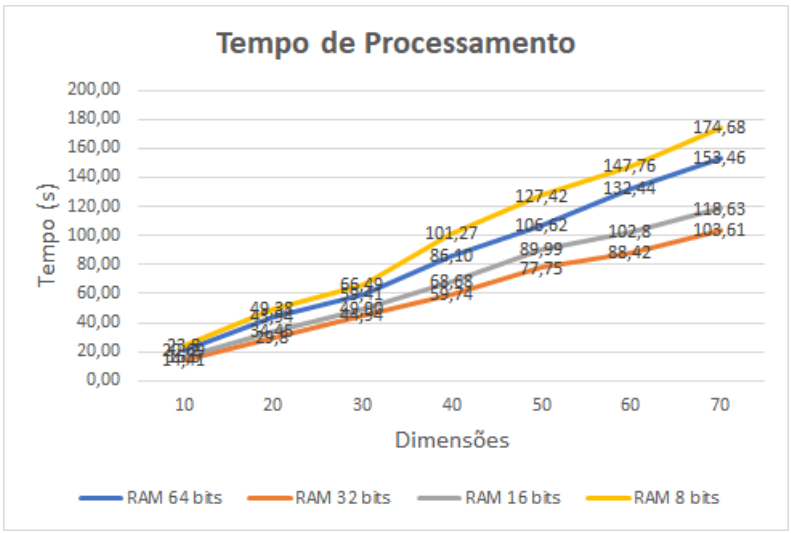

Figura 4.4: Tempo de processamento relativo à classificação por partido baseado no corpo da notícia

4.4.2 Classificação Baseada na Manchete

A manchete da notícia é a parte de destaque da mensagem, onde o autor do artigo realiza a chamada para os leitores. Costuma ser composto de frases pequenas e atrativas, e revela o assunto principal. Com o intuito de se chamar a atenção para o conteúdo, geralmente são utilizados termos de destaque ou até mesmo apelativos, tudo isso para que desperte o interesse de interlocutor pela a notícia. Em comparação à classificação baseada no texto da notícia, a classificação baseada na manchete da notícia é uma tarefa mais complexa para a classificação, devido a pouca quantidade de termos para descrever ideia da notícia. Para descrever por quantitativos, foram utilizados como índice 2161 termos distintos, em 1147 documentos.

Foram realizadas avaliações utilizando a redução de dimensionalidade através da SVD para variações de 10 até a dimensão de tamanho 90. Após a binarização das dimensões reduzidas, para cada dimensão, foi avaliado o WiSARD para 8,16,32 e 64 bits de memória. Foram rotuladas as notícias de acordo com suas fontes, através de suas siglas.

Observamos que nas Figuras 4.5 e 4.6, a acurácia de cerca de 75% do modelo utilizando 64 bits de endereçamento decai ao se reduzir a quantidade dos bits de endereçamento de memória. Este fato explica o aumento do potencial de generalização do modelo relacionado ao tamanho da memória. Com relação à quantidade de dimensões, observamos um aumento da acurácia relacionado às dimensões que identificamos como ideais para o modelo.

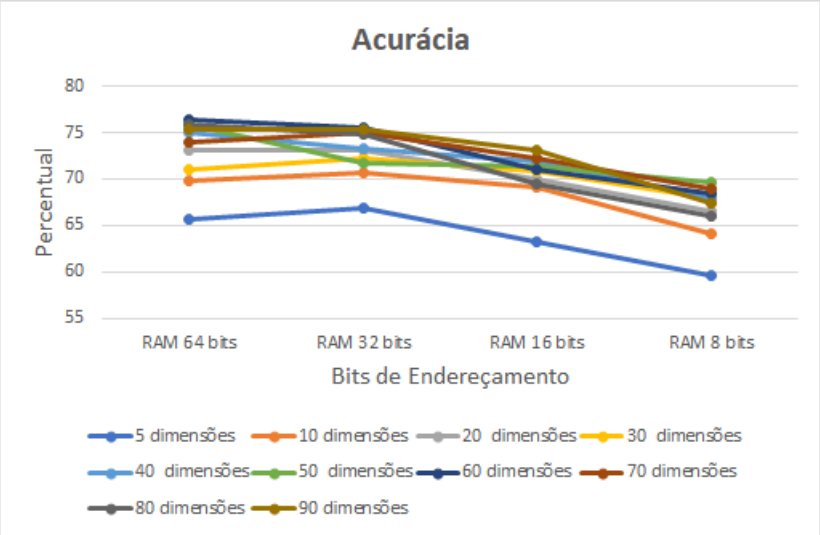

Figura 4.5: Acurácia relativa à classificação por polaridade baseada na manchete da notícia

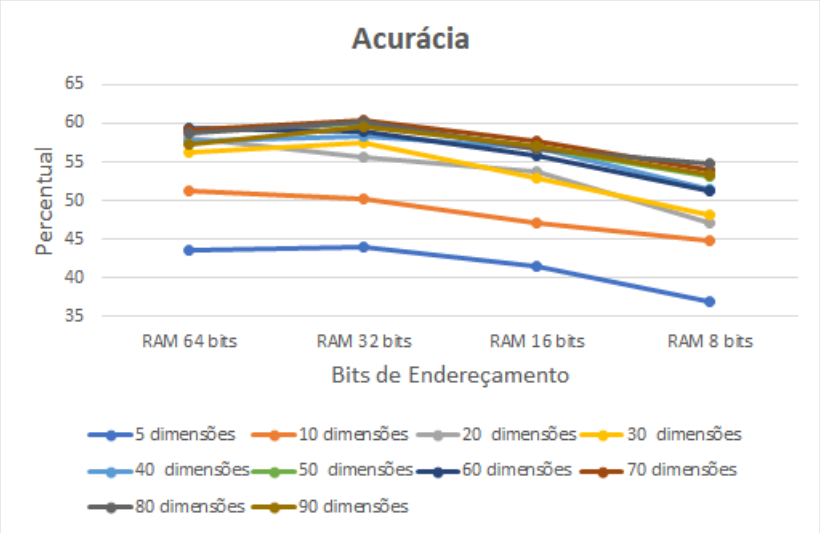

Figura 4.6: Acurácia relativa à classificação por partido baseada na manchete da notícia

Nas Figuras 4.7 e 4.8, observamos que o tempo de processamento cresce de

forma linear de acordo com a quantidade de dimensões dos dados utilizados no WiSARD.

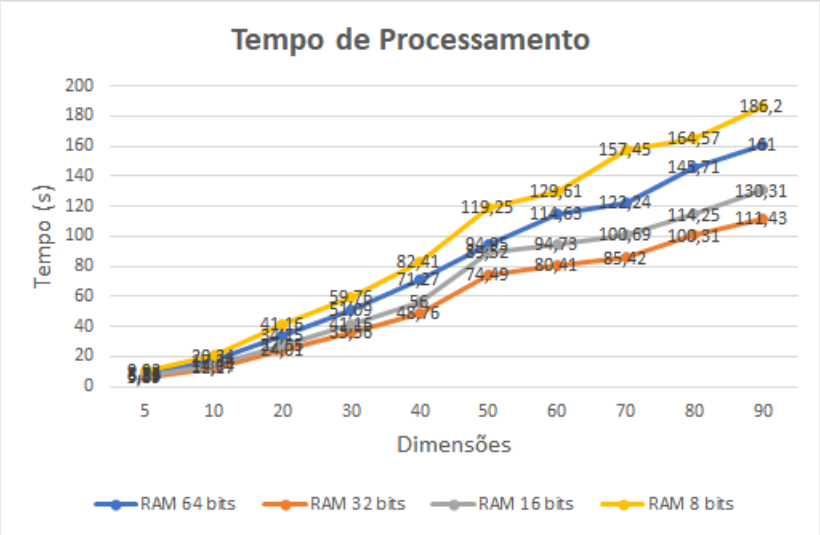

Figura 4.7: Tempo de processamento relativo à classificação por polaridade baseado na manchete da notícia

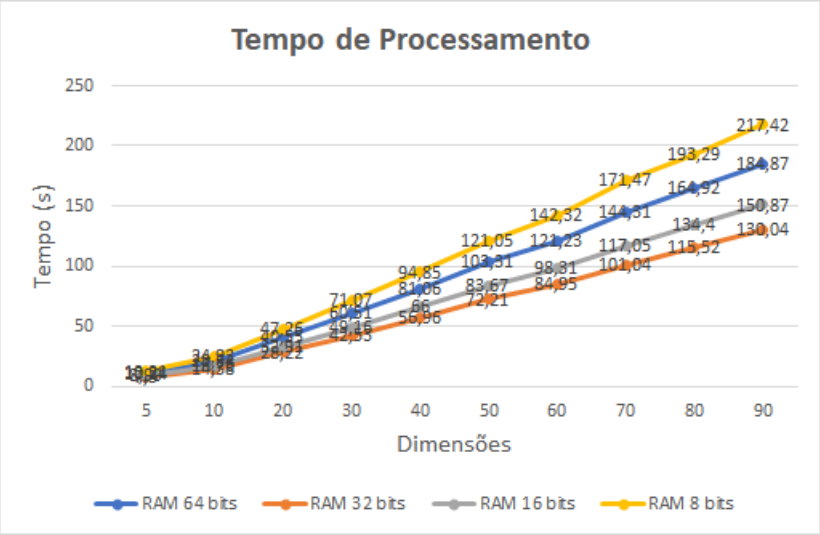

Figura 4.8: Tempo de processamento relativo à classificação por partido baseado na manchete da notícia

4.5 AVALIAÇÃO DA EVOLUÇÃO TEMPORAL DA POLARIDADE

O objetivo deste experimento é avaliar, dentro do conjunto de notícias coletados de um noticiário amplamente conhecido, a evolução temporal das notícias de acordo com o resultado do classificador. Foram consideradas que as notícias possuem um viés político em sua elaboração.

Como base de dados do noticiário a ser avaliado, foram coletadas notícias dos anos de 2016 e 2017 da página do Jornal Nacional (http://g1.globo.com/jornal-nacional/). É importante ressaltar que, para a avaliação temporal, os arquivos foram organizados de acordo com sua ordem de publicação, incluindo-se o ano e o mês no nome do arquivo. Como existem diversos temas abordados pelo referido veículo de mídia, em vez de utilizarmos todas as notícias coletadas, selecionamos notícias sobre alguns personagens políticos deste período. Para selecionar os personagens, filtramos as notícias de acordo com com as palavras-chave 'Lula', 'Dilma', 'Temer' e 'Moro', utilizadas na manchete.

Executando todas as etapas do processo de mineração de textos para realização do experimento, foram utilizados os corpos das notícias coletadas como dados, passando pela fase de pré-processamento, ponderação dos termos, redução de dimensionalidade e binarização dos vetores. Para a classificação, foi utilizado o algoritmo ClusWiSARD [14] com o conjunto de treinamento rotulado abordado na Seção 4.2.

Os dados coletados da página do Jornal Nacional não receberam nenhum rótulo e foram apresentados em ordem temporal. Após cada dado apresentado, de acordo a classificação obtida, este dado era incluído no conjunto de treinamento, pertencendo à classe atrubuída pela ClusWiSARD. Como parâmetros da ClusWiSARD, utilizamos *clusters* de tamanho 40, similaridade mínima de 50% e tamanho de endereçamento de memória de 32 bits.

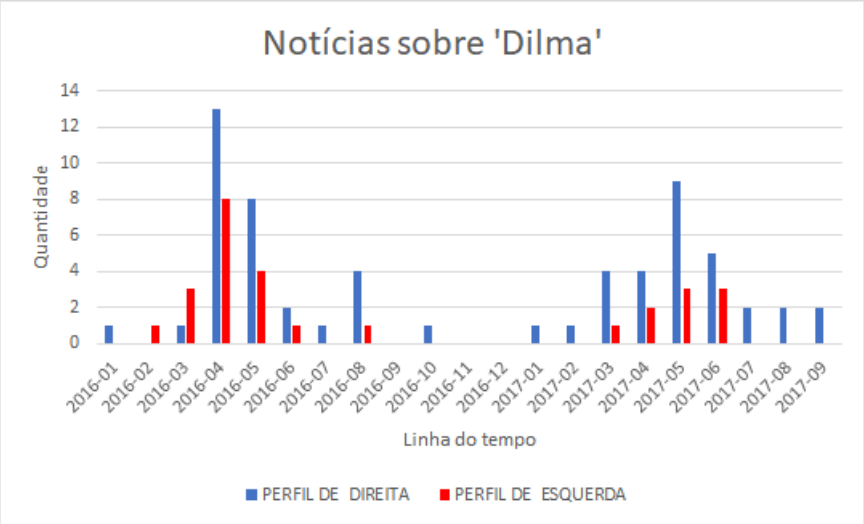

Figura 4.9: Distribuição temporal de notícias com a palavra-chave 'Dilma' por polaridade

Após a classificação das notícias, é possível extrair algumas conclusões de acordo com fatos políticos ocorridos nos anos de 2016 e 2017. Na Figura 4.9, podemos observar o grande volume de notícias entre abril e agosto de 2016, período em que houve a votação da Câmara dos Deputados sobre o *impeachment* da ex-presidente Dilma e a assunção do presidente Temer. O segundo período de grande volume de notícias é relacionado ao julgamento da chapa Dilma-Temer e delações que envolveram a ex-presidente.

Na Figura 4.10, que é referente à notícias relativas ao termo 'Temer', podemos observar um aumento na quantidade de notícias em maio de 2017, mês em que houve, no dia 18, a divulgação pela imprensa um áudio sobre uma conversa entre o presidente e um empresário sobre supostos atos de corrupção. Nesse período também estava ocorrendo o julgamento da chapa Dilma-Temer que se iniciou em abril e terminou em junho. Este alto volume de notícias se deu até agosto de 2017, quando, no dia 2, houve o arquivamento do processo contra o presidente Michel Temer por corrupção passiva pela Câmara dos Deputados. Nos meses de julho e

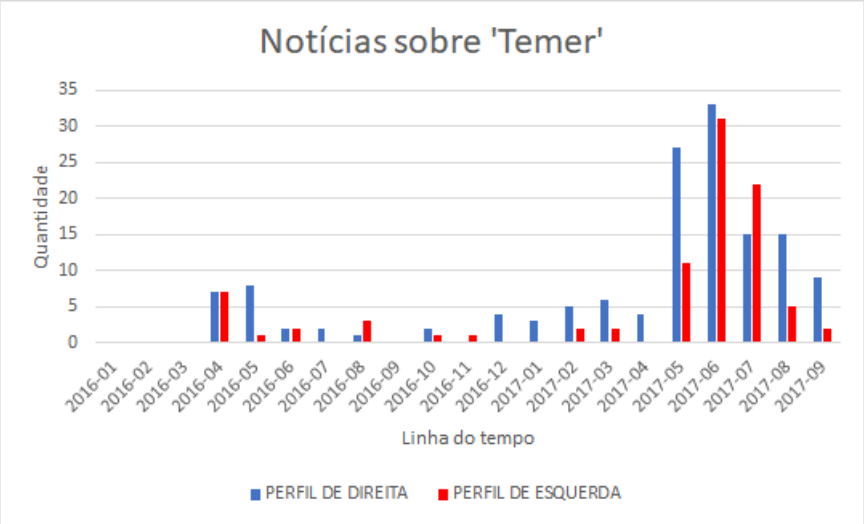

Figura 4.10: Distribuição temporal de notícias com a palavra-chave 'Temer' por polaridade

agosto, é possível observar uma transição no perfil das notícias referentes ao termo.

Com relação às notícias referentes ao ex-presidente Lula, observamos que o volume de notícias um pouco maior antes da votação pela Câmara dos Deputados pelo *impeachment* da ex-presidente Dilma. Este volume é relacionado à nomeação do ex-presidente Lula como ministro da Casa Civil e o próximo grande volume de notícias coincide com o volume da Figura 4.12 e é referente aos depoimentos feitos pelo ex-presidente Lula ao Juiz Sérgio Moro e notícias relacionadas à delações.

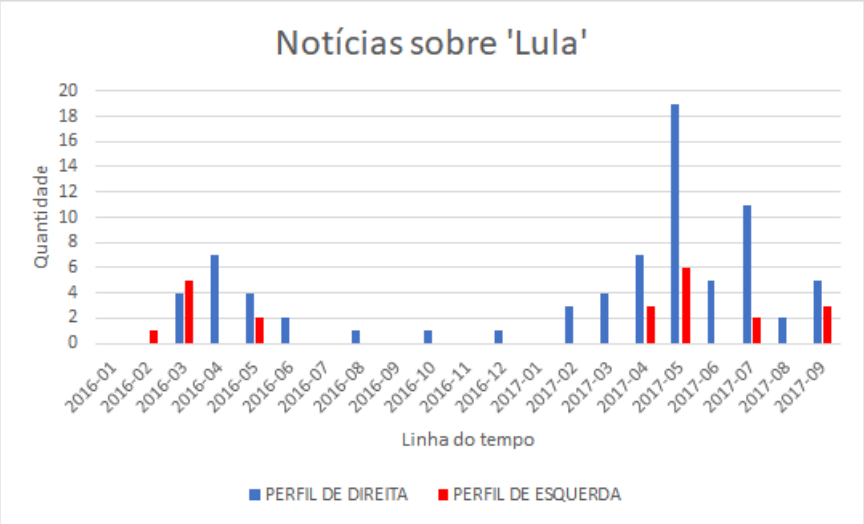

Figura 4.11: Distribuição temporal de notícias com a palavra-chave 'Lula' por polaridade

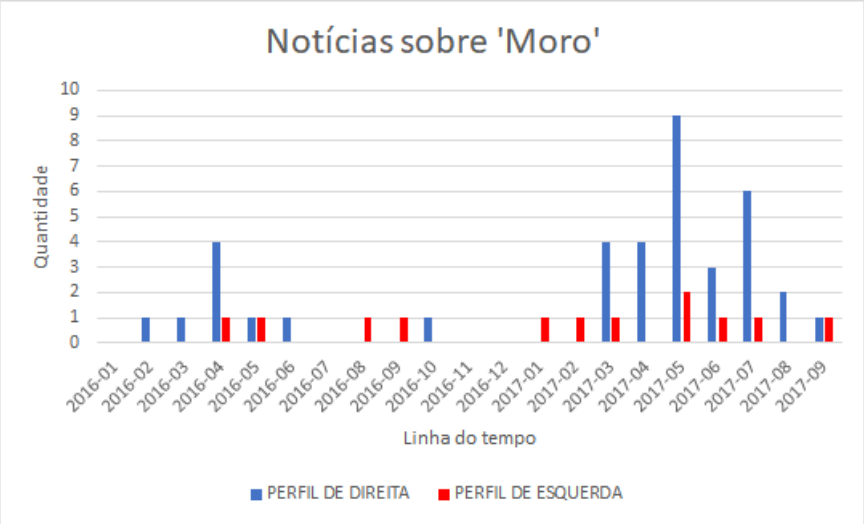

Figura 4.12: Distribuição temporal de notícias com a palavra-chave 'Moro' por polaridade

5 CONCLUSÃO

O problema da identificação de polaridade em notícias é uma aplicação interessante para a aprendizagem por máquinas. Este trabalho mostrou todas a etapas do processo de descoberta do conhecimento, discriminando suas peculiaridades e, como classificador, a Rede Neural sem Peso WiSARD. Foi abordada a classificação de notícias políticas em português por partido e buscou-se identificar a polaridade das notícias de acordo com a posição partidária atual no Brasil. A capacidade de aprendizado *on-line* do WiSARD e da ClusWiSARD, permitindo a inclusão de novos elementos em seu conjunto de treinamento sem a necessidade de se processar todos os dados, não é característica usual dos métodos de classificação e aglomeração. Isso as torna boas opções para utilização em identificação de polaridade e em classificação de notícias. A redução de dimensionalidade através da decomposição em valores singulares proporciona ao processo de mineração de textos, um ganho temporal considerável, proporcionando pouca redução nas medidas de acurácia ao mesmo tempo reduzir consideravelmente a dimensão dos dados textuais.

Conforme pode ser observado no comparativo entre os classificadores, a RNSP WiSARD foi comparada à outros algoritmos amplamente conhecidos no processo de mineração de textos, obtendo desempenho comparável à melhor alternativa encontrada, que foi a Regressão Logística e desempenho superior aos classificadores SVM, *Naive Bayes* e *Gradient Tree Boosting*. Por este motivo, uma vez feitas as comparações iniciais, a WiSARD (ou uma de suas variantes, a ClusWiSARD) foi escolhida para a realização de experimentos sobre aspectos não previamente analisados na literatura como, por exemplo, a evolução temporal da tendência das notícias.

O WiSARD, originalmente desenvolvido para reconhecimento de imagens, se

portou de forma altamente satisfatória com a conversão dos dados originalmente textuais, fato que demonstra a importância do mapeamento para a representação binária juntamente com a capacidade desta rede neural em trabalhar com vetores de dados reais em mineração de textos ou dados. Para resolução de outros tipos de problemas, pode-se investir uma dedicação maior ao modelo de representação, em vez de se buscar um desempenho melhor em outros classificadores. Apresentamos também resultados sobre a evolução temporal da polaridade de notícias, levando em conta a classificação não supervisionada das mesmas com a ClusWiSARD.

Os resultados demonstraram que a utilização do corpo da notícia no processo de mineração de textos obtiveram melhores resultados em acurácia. A utilização das manchetes das notícias se mostrou como mais uma alternativa para identificar a polaridade em notícias, por apresentar uma quantidade menor de termos e alcançar um nível de acurácia comparável à utilização do corpo da notícia. Sua utilização pode ser considerada ao se construir uma ferramenta para avaliação de polaridade de notícias compartilhadas em redes sociais, onde normalmente não se tem disponível de imediato o corpo da notícia.

5.1 TRABALHOS FUTUROS

O desenvolvimento deste trabalho mostrou que a identificação de polaridade em notícias através da mineração de textos é um processo muito complexo e multidisciplinar, fazendo com que o tempo seja curto para explorar todas a técnicas existentes. Durante o desenvolvimento, observaram-se algumas possibilidades de melhorias na metodologia adotada. Abaixo estão algumas sugestões que podem ser implementadas e testadas em trabalhos futuros:

- **Notícias imparciais** - O modelo utilizado neste trabalho aborda a classi-

ficação de notícias avaliando um viés ideológico de PT/PSOL ou PMDB/PSDB. Seria um bom campo a ser explorado a pesquisa para identificação de notícias imparciais, ou seja, identificar se uma notícia possui ou não um viés.

- **Exploração da avaliação temporal** - A utilização da ClusWiSARD neste trabalho se mostrou satisfatória para avaliação temporal das notícias em uma janela de tempo de cerca de 2 anos. Um campo a ser explorado, seria a classificação de notícias com a utilização da janela de tempo com esquecimento, visto que um partido político pode mudar sua posição sobre um determinado assunto com o passar do tempo.

- **Entidades nomeadas** - A inclusão de reconhecimento de entidades nomeadas irá aumentar a eficiência do modelo, visto que no modelo atual o processamento das notícias é realizado sem a identificação de nomes próprios, empresas, órgãos governamentais entre outras entidades. O reconhecimento dessas entidades pode reduzir os erros de classificação do modelo.

- **Novos experimentos** - Outra opção para continuidade do trabalho é a abordagem de outros temas de notícias como por exemplo notícias de conteúdo religioso, notícias criminais etc. Também podem ser realizados novos experimentos com grandes volumes de documentos, para avaliar o desempenho da metodologia com conjuntos de dados maiores;

- **Interface** - Pode ser realizada a construção de uma interface, utilizando o aprendizado *on-line*, que realize a avaliação de polaridade em uma notícia através do *browser* ou que avalie um notícia compartilhada em uma rede social, proporcionando ao usuário um suporte à sua avaliação crítica de forma imediata.

REFERÊNCIAS

[1] ALEKSANDER, I. From WISARD to MAGNUS: a family of weightless virtual neural machines. **RAM-Based Neural Networks**, Canterbury, v.9, 1998.

[2] ALEKSANDER, I.; DE GREGORIO, M.; FRANÇA, F. M. G.; LIMA, P. M. V.; MORTON, H. A brief introduction to Weightless Neural Systems. In: ESANN, Bruges, Bélgica. **Anais...** UCL/ELEN, 2009. p.299–305.

[3] ALEKSANDER, I.; THOMAS, W.; BOWDEN, P. WISARD: a radical step forward in image recognition. **Sensor review**, Bingley, Reino Unido, v.4, n.3, p.120–124, 1984.

[4] AMIN, M. T. A.; AGGARWAL, C.; YAO, S.; ABDELZAHER, T.; KAPLAN, L. **Unveiling polarization in social networks**: a matrix factorization approach. Champaign, Illinois, USA: IEEE, 2017.

[5] AN, J.; KWAK, H. What Gets Media Attention and How Media Attention Evolves Over Time-Large-scale Empirical Evidence from 196 Countries. **arXiv preprint arXiv:1704.01425**, Stanford, California, 2017.

[6] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação**: conceitos e tecnologia das máquinas de busca. Rio de Janeiro, Brasil: Bookman Editora, 2013.

[7] BAKSHY, E.; MESSING, S.; ADAMIC, L. A. Exposure to ideologically diverse news and opinion on Facebook. **Science**, Washington, D.C., USA, v.348, n.6239, p.1130–1132, 2015.

[8] BALASUBRAMANYAN, R.; COHEN, W. W.; PIERCE, D.; REDLAWSK, D. P. Modeling polarizing topics: when do different political communities res-

pond differently to the same news? In: ICWSM, Dublin, Irlanda. **Anais...** NUI Galway, 2012.

[9] BANDEIRA, L. C. **NC-WISARD**: uma interpretação sem pesos do modelo neural neocognitron. 2010. Tese (Doutorado em Ciência da Computação) — Universidade Federal do Rio de Janeiro.

[10] BLEDSOE, W. W.; BROWNING, I. Pattern recognition and reading by machine. In: PAPERS PRESENTED AT THE DECEMBER 1-3, 1959, EASTERN JOINT IRE-AIEE-ACM COMPUTER CONFERENCE, Boston, Massachusetts. **Anais...** ACM, 1959. p.225–232.

[11] BRADFORD, R. B. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In: ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 17. **Proceedings...** ACM, 2008. p.153–162.

[12] BRASIL, B. **De criação na ditadura até o rompimento com governo**: o pmdb em 10 capítulos. Acesso em: 2017-10-05, Disponível em:<http://www.bbc.com/portuguese/noticias/2016/03/160330_pmdb_historia_ms_ss>.

[13] BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD WORKSHOP: LANGUAGES FOR DATA MINING AND MACHINE LEARNING. **Anais...** arXiv:1309.0238, 2013. p.108–122.

[14] CARDOSO, D. d. O.; CARVALHO, D. S.; ALVES, D. S.; SOUZA, D. F. P. de; CARNEIRO, H. C.; PEDREIRA, C. E.; LIMA, P. M.; FRANÇA, F. M. Credit analysis with a clustering RAM-based neural classifier. In: ESANN, Bruges, Bélgica. **Anais...** UCL/ELEN, 2014.

[15] CARNEIRO, H. C.; FRANÇA, F. M.; LIMA, P. M. Multilingual part-of-speech

tagging with weightless neural networks. **Neural Networks**, Elsevier, v.66, p.11–21, 2015.

[16] CAVALCANTI, R. D.; LIMA, P. M. V.; GREGORIO, M. D.; MENASCHE, D. S. Evaluating weightless neural networks for bias identification on news. In: IEEE 14TH INTERNATIONAL CONFERENCE ON NETWORKING, SENSING AND CONTROL (ICNSC), 2017., Calabria, Italy. **Anais...** IEEE, 2017. p.257–262.

[17] CHEN, H. **Knowledge management systems**: a text mining perspective. Arizona, USA: Knowledge Computing Corporation, 2001.

[18] DEERWESTER, S. Improving information retrieval with latent semantic indexing. **Journal of the American Society for Information Science**, New York, USA, 1988.

[19] DIAP, A. **Eleições 2014**: direita política se populariza no brasil. Acesso em: 2017-10-05, Disponível em:<https://tinyurl.com/y9q772hb>.

[20] DOSCIATTI, M. M.; FERREIRA, L. P. C.; PARAISO, E. C. Identificando emoçoes em textos em português do brasil usando máquina de vetores de suporte em soluçao multiclasse. **ENIAC-Encontro Nacional de Inteligência Artificial e Computacional**, Fortaleza, Ceará, Brasil, 2013.

[21] EBECKEN, N. F.; LOPES, M. C. S.; COSTA, M. C. et al. Mineração de textos. **Sistemas inteligentes: fundamentos e aplicações**, São Carlos, p.337–370, 2003.

[22] FAN, R.-E.; CHANG, K.-W.; HSIEH, C.-J.; WANG, X.-R.; LIN, C.-J. LIBLINEAR: a library for large linear classification. **Journal of machine learning research**, USA, v.9, n.Aug, p.1871–1874, 2008.

[23] FIGUEIRA, C. V. **Modelos de regressão logística**. 2006. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

[24] GOMES, H. J. C. **Text Mining**: análise de sentimentos na classificação de notícias. 2013. Tese (Doutorado em Ciência da Computação) — Universidade Nova de Lisboa, Lisboa, Portugal.

[25] GRANJA, B. **Geração só Manchete**. Acesso em: 2017-09-20, Disponível em:<https://tinyurl.com/yd4wwdwh>.

[26] GRIECO, B. P.; LIMA, P. M.; DE GREGORIO, M.; FRANÇA, F. M. Producing pattern examples from "mental" images. **Neurocomputing**, Elsevier, v.73, n.7, p.1057–1064, 2010.

[27] GUERRA, P. C.; MEIRA JR, W.; CARDIE, C. Sentiment analysis on evolving social streams: how self-report imbalances can help. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, 7. **Proceedings...** ACM, 2014. p.443–452.

[28] HAYKIN, S.; NETWORK, N. A Comprehensive Foundation. **Neural Networks**, Ontario, Canada, v.2, n.2004, p.41, 2004.

[29] HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. New York: John Wiley & Sons, 2013. v.398.

[30] HUNTER, J. D. Matplotlib: a 2d graphics environment. **Computing In Science & Engineering**, IEEE Computing in Science Engineering, v.9, n.3, p.90–95, 2007.

[31] KOENIG, M. E. Information driven management: the new, but little perceived, business zeitgeist. **Libri**, USA, v.50, n.3, p.174–190, 2000.

[32] LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. **Discourse Processes**, Reino Unido, v.25, n.2-3, p.259–284, 1998.

[33] LEWIS, D. D.; RINGUETTE, M. A comparison of two learning algorithms for text categorization. In: THIRD ANNUAL SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL, Las Vegas, Nevada. **Anais...** University of Nevada, 1994. v.33, p.81–93.

[34] LOTAN, G. **Obama nails why the political climate is so polarized**. Acesso em: 2016-11-25, Disponível em:<https://boingboing.net/2016/11/25/beyond-fake-news-the-constr.html>.

[35] MARTINAZZO, B. **Um Método de Identificação de Emoções em Textos Curtos para o Português do Brasil**. 2010. Dissertação (Mestrado em Ciência da Computação) — PUC-PR, Paraná, Brasil.

[36] MAYRINK, V. T. d. M. **Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo**. 2015. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Juiz de Fora.

[37] MCCALLUM, A.; NIGAM, K. A comparison of event models for Naive Bayes text classification. In: IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION. **Anais...** AAAI Press, 1998. v.752, p.41–48.

[38] MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Kluwer Academic Publishers, v.5, n.4, p.115–133, 1943.

[39] MITCHELL, T. M. et al. **Machine learning**. Boston, MA: McGraw-Hill, 1997.

[40] OLIVEIRA CARDOSO, D. de. **Uma arquitetura para agrupamento de dados em fluxo contínuo baseada em redes neurais sem pesos**. 2012.

Tese (Doutorado em Ciência da Computação) — Universidade Federal do Rio de Janeiro.

[41] OLIVEIRA CARDOSO, D. de. **Rejection-oriented learning without complete class information**. 2017. Tese (Doutorado em Ciência da Computação) — Universidade Federal do Rio de Janeiro.

[42] PÁDUA BRAGA, A. de; LEON FERREIRA, A. C. P. de; LUDERMIR, T. B. **Redes neurais artificiais**: teoria e aplicações. Rio de Janeiro, Brasil: LTC Editora, 2014.

[43] PEDREGOSA, F. et al. Scikit-learn: machine learning in python. **Journal of Machine Learning Research**, USA, v.12, n.Oct, p.2825–2830, 2011.

[44] PONTIL, M.; VERRI, A. Properties of support vector machines. **Neural Computation**, Massachusetts, USA, v.10, n.4, p.955–974, 1998.

[45] PRADO, C. B.; FRANÇA, F. M.; DIACOVO, R.; LIMA, P. M. The Influence of Order on a Large Bag of Words. In: INTELLIGENT SYSTEMS DESIGN AND APPLICATIONS, 2008. ISDA'08. EIGHTH INTERNATIONAL CONFERENCE ON, Kaohsiung, Taiwan. **Anais...** IEEE, 2008. v.1, p.432–436.

[46] PULLELLA, P. **Pope denounces virus of polarization**. Acesso em: 2016-11-25, Disponível em:<https://tinyurl.com/jowg66k>.

[47] RANGEL, F.; FIRMINO, F. **Python Weightless Artificial Neural Network**. Acesso em: 2016-01-20, Disponível em:<https://github.com/firmino/PyWANN.

[48] REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: **Encyclopedia of database systems**. New York, USA: Springer, 2009. p.532–538.

[49] RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Why should i trust you?: explaining the predictions of any classifier. In: ACM SIGKDD INTERNATIONAL

CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., San Francisco, California, USA. **Proceedings...** ACM, 2016. p.1135–1144.

[50] ROCHA, V. da. **Saiba como surgiu e o que defende cada partido**. Acesso em: 2017-10-05, Disponível em:<https://tinyurl.com/yaa3wy2l>.

[51] SAMUELS, D.; ZUCCO, C. The power of partisanship in Brazil: evidence from survey experiments. **American Journal of Political Science**, Michigan, v.58, n.1, p.212–225, 2014.

[52] STAFF, R. **Danish parties agree on tougher border controls**. Acesso em: 2017-10-05, Disponível em:<https://tinyurl.com/y6wh329a>.

[53] SUYKENS, J. A.; VANDEWALLE, J. Least squares support vector machine classifiers. **Neural processing letters**, Kluwer Academic Publishers, v.9, n.3, p.293–300, 1999.

[54] TAROUCO, G. d. S.; MADEIRA, R. M. Partidos, programas e o debate sobre esquerda e direita no Brasil. **Revista de Sociologia e Política**, Curitiba, v.21, p.149 – 165, 03 2013.

[55] THIAGO GUIMARãES, R. S. e. **Como as eleições municipais desidratam os partidos de esquerda**. Acesso em: 2017-10-05, Disponível em:<http://www.bbc.com/portuguese/brasil-37710397>.

[56] VASCONCELLOS, F. **Maioria dos partidos se posiciona como de centro. Veja quem sobra no campo da direita e da esquerda**. Acesso em: 2017-10-05, Disponível em:<https://tinyurl.com/y72rvf2r>.

[57] WALT, S. v. d.; COLBERT, S. C.; VAROQUAUX, G. The NumPy array: a structure for efficient numerical computation. **Computing in Science & Engineering**, arXiv:1102.1523v1, v.13, n.2, p.22–30, 2011.

[58] WIKIPEDIA. **Radikale Venstre**. Acesso em: 2017-10-05, Disponível em:<https://da.wikipedia.org/wiki/Radikale_Venstre>.

[59] WINTER, B. **Exclusive**: brazil opposition leader will seek economic reforms. Acesso em: 2017-10-05, Disponível em:<https://tinyurl.com/y9w9n7gf>.

APÊNDICE A TERMINOLOGIA

Em nosso trabalho, consideramos alguns termos que podem ser confundidos com outros significados. Nesta seção, apresentamos algumas definições do que consideramos neste trabalho e definições de termos que não estão no escopo.

- **Notícia tendenciosa** - Quando uma notícia contém uma opinião implícita ou explícita, em outras palavras, ela apresenta os fatos de acordo com seu ponto de vista, de forma parcial;
- **Polaridade** - Nos referimos às opiniões opostas sobre um determinado assunto. Consideramos o conceito de forma distinta da **análise de sentimentos**, visto que notícias com a mesma polaridade podem ter sentimentos positivos e negativos. Por exemplo, as frases 'Temer encaminhou a reforma.' e 'A reforma ainda não foi votada.' apresentam mesma polaridade e sentimentos distintos;
- **Transparência da notícia** - Esclarecimento quanto à existência de uma opinião polarizada em uma notícia;
- **Notícias falsas (*fake news*)** - Refere-se à notícias fabricadas, com a intenção de enganar. O conteúdo da notícia é distorcido e inventado. Não faz parte do escopo deste trabalho. Consideramos que as notícias utilizadas se referem a fatos verídicos e são de fontes confiáveis;
- **Detecção de autoria** - Identifica a assinatura do único autor, de acordo com seu estilo de escrita. Também não está no escopo deste trabalho, pois um editorial contém diversos autores e a intenção do trabalho é identificar uma opinião implícita;

- **Detecção de veículo** - Identifica a assinatura do veículo de mídia (e.g. jornal, revista, site) conjunta de seus autores, de acordo com seu estilo de escrita, polaridade, etc. Neste trabalho, consideramos autores por natureza polarizados, tendo em vista que as notícias utilizadas são das páginas dos partidos. Dessa forma, mapeamos o problema de detecção de polaridade dentro da detecção de veículo; e
- **Rumor** - Notícia duvidosa, que é questionada. Consideramos as notícias verídicas e a identificação rumores não se enquadram no escopo do trabalho.

APÊNDICE B REDES NEURAIS WISARD E CLUSWISARD

B.1 REDE NEURAL SEM PESO WISARD

O sistema WiSARD (Wilkie, Stonham & Aleksander's Recognition Device) é um sistema com diversos discriminadores trabalhando juntos, cada discriminador com diversos neurônios RAM. Na WiSARD, cada discriminador é responsável pelo reconhecimento de uma única classe. A WiSARD foi a primeira rede neural artificial a ser patenteada e produzida comercialmente, sendo também o modelo de rede neural sem peso mais representativo [26].

Originalmente, o WiSARD foi projetada para o reconhecimento de imagens, em preto e branco através de uma implementação em hardware. Por conta dessa peculiaridade, o formato de entrada dos dados é em sequência binária. Essa sequência também é denominada retina. Cada discriminador é composto por neurônios RAM, de tamanho definido pelo usuário. A escolha do tamanho das tuplas da memória RAM influenciam diretamente na quantidade de neurônios que serão utilizados na rede. Caso seja necessário trabalhar com valores diferentes de binários, é necessário realizar um tratamento para representação em cadeia de bits [40].

Os padrões apresentados à rede são separado em tuplas. Cada uma dessas tuplas estão relacionadas com uma memória RAM e o conhecimento estará armazenado nessas memórias de acordo com os endereços ativados. As tuplas do mapeamento do padrão de são utilizadas para armazenar um padrão na memória RAM e também são utilizadas para recuperar os valores armazenados, por ocasião da classificação.

[1].

B.1.1 Treinamento

Na primeira etapa do treinamento, denominada inicialização, o tamanho da retina do padrão de entrada e a quantidade de k bits de memória definem a quantidade de memórias RAM. Em seguida, o conteúdo dos endereços de cada memória é setado com 0. A partir de então, é gerado um mapeamento pseudo-aleatório para os padrões que serão apresentados. cada tupla de tamanho k é associada a uma RAM específica. Na Figura B.1, podemos observar que os bits da retina são mapeados para as n-tuplas, no caso da figura B.1, n=3.

Na segunda etapa do treinamento, cada padrão é apresentado ao discriminador de sua respectiva classe. De acordo com o mapeamento definido na primeira etapa, cada endereço ativado por cada tupla é setado com 1 em sua respectiva posição. O mecanismo pode ser observado na Figura B.1, com a formação de um 'T'.

B.1.2 Classificação

O reconhecimento de um novo padrão é feito de forma parecida com o treinamento, com a diferença que o padrão é apresentado a todos os discriminadores. Ao apresentar o novo padrão a cada discriminador, o conteúdo dos endereços de memória ativados são recuperados e somados, obtendo-se a resposta r de cada discriminador, que é equivalente ao número de RAMs com saída 1. O valor de r atinge seu máximo X o padrão apresentado pertencer ao conjunto de treinamento. Outros valores de r são uma podem ser considerados como tipo de 'medida de similaridade'

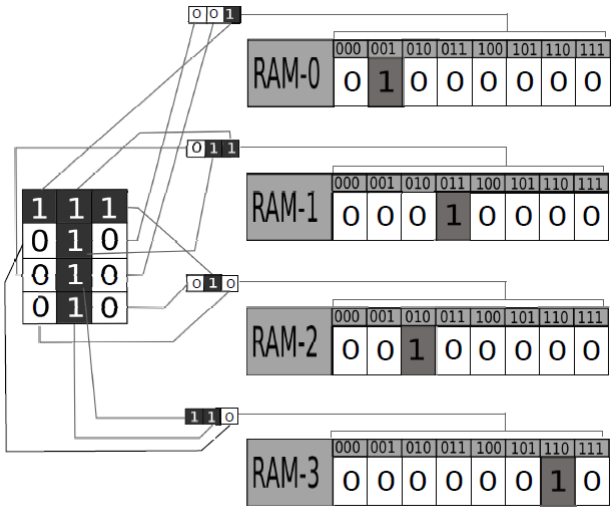

Figura B.1: Treinamento do caractere 'T' [15]

do padrão de apresentado em relação aos padrões no conjunto de treinamento [2].

A Figura B.2 ilustra o reconhecimento dos discriminadores de cada letra do alfabeto para o padrão apresentado. Para cada discriminador, o valor de r equivale ao seu respectivo placar.

Durante a classificação, é utilizado um dispositivo chamado calculador que identifica a maior resposta r e segundo [42], e também realiza as atividades:

- medir a confiança absoluta, em outras palavras, informar qual a classe que foi reconhecida ou o quanto se parece com as classes treinadas;

- medir a confiança relativa, que é realizada através da relação $C = D/R_jmax$, onde C é a confiança relativa, D é a diferença entre as duas maiores respostas e R_jmax é a maior resposta.

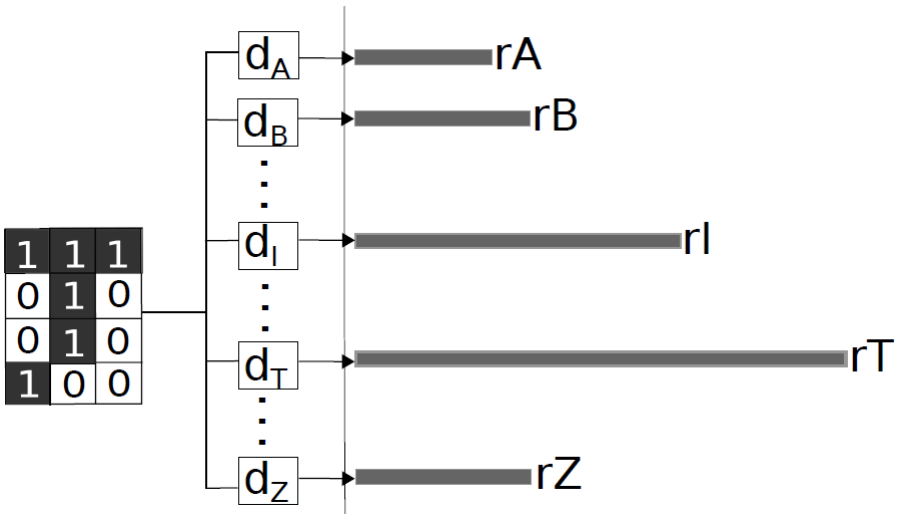

Figura B.2: Modelo do Classificador WiSARD e o placar dos discriminadores [15]

B.1.3 O Mecanismo de *Bleaching*

Para cada padrão apresentado a um discriminador, a posição de memória reconhecida é marcada com 1. Com o aumento de padrões apresentados, variações no padrão e ruídos vão sendo adicionados aos neurônios. Uma quantidade grande de padrões de treinamento pode fazer com que a rede fique saturada, gerando dificuldades para a classificar um novo padrão.

Uma solução para este problema foi proposta por [26] e se chama *bleaching*, onde o conteúdo da memória é incrementado de 1 a cada padrão apresentado. Ao final do treinamento, os valores dos conteúdos da memória irão variar de 0 até a quantidade de padrões de treinamento treinada. Quando ocorre o empate na classificação entre os discriminadores, o valor do *bleaching* é utilizado como um limiar para não contabilizar os neurônios que tiveram resposta abaixo deste limiar. Em caso de novo empate, este valor é incrementando de 1 em 1 até que seja encontrada a classe correta [9].

B.2 CLUSWISARD

Como na WiSARD cada discriminador é responsável pelo reconhecimento de uma única classe, padrões muito diferentes de um mesma classe podem ser treinados. Essa situação pode ocasionar classificações incorretas de padrões apresentados posteriormente. A ClusWiSARD busca resolver este problema através da utilização de agrupamentos dos padrões de entrada no processo de treinamento. O uso de discriminadores como representantes de *clusters* é a principal diferença disso [14].

O principal diferencial da ClusWiSARD se trata na forma do armazenamento do conhecimento. Diferente da WiSARD, utiliza um grupo de discriminadores por classe, conseguindo identificar subpadrões. No processo de treinamento, os padrões são apresentados à rede e, de acordo com a resposta ao padrão apresentado aos discriminadores existentes, o padrão é aprendido no caso da resposta ser maior que o limiar de aceitação. No caso aa resposta menor que o limiar, é criado um novo discriminador, que aprende o padrão. O limite de aceitação é proporcional ao número de elementos no *cluster* [14].

O processo de classificação da ClusWiSARD é muito parecido com o da WiSARD, onde um padrão novo é apresentado a rede e os discrimindadores enviam sua resposta. Em caso de dois discriminadores de classes distintas empatarem, o processo de *bleaching* é utilizado. A classe associada ao discriminador com maior resposta é escolhida [41].

APÊNDICE C CLASSIFICADORES E MÉTRICAS DE AVALIAÇÃO

C.1 CLASSIFICADORES

C.1.1 Regressão Logística

Em situações onde necessitamos obter uma resposta qualitativa (discreta), podemos utilizar a técnica de regressão logística para calcular ou prever a probabilidade da classe de um documento específico [23]. Neste modelo, assumimos que a variável dependente seja binária. Esta variável assume dois valores, normalmente, 0 e 1, respectivamente nomeados, neste trabalho, como "Classe 1"e "Classe 2". No nosso caso, o evento de interesse é conhecer a classificação de um determinado texto.

Seguindo a notação de [29], dado um conjunto de p variáveis explicativas independentes, nós organizamos os pontos de dados fornecidos n na matriz $\mathcal{X}$,

$$\mathcal{X} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & & & \\ x_{i0} & x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & & & & \\ x_{n0} & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \tag{C.1}$$

onde $x_i^T = (x_{i0}, x_{i1}, x_{i2}, ..., x_{ip})$ é o i-ésimo vetor, igual à i-ésima linha da matriz X, para $i = 1, 2, \dots, n$ e $j = 0, 1, \dots, p$. Nós assumimos que $x_{i0} = 1$ para todo i.

Denominamos por $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_p)^T$ o vetor de parâmetros desconhecidos, onde $\tilde{\beta}_j$ é o j-ésimo parâmetro, associado à j-ésima variável explicativa.

Deixamos X_i e Y_i sendo duas variáveis aleatórias que denotam o vetor de características (variáveis explicativas independentes) e o resultado (variável dependente). No modelo de regressão logística múltipla, a probabilidade de sucesso é dada por

$$\pi_i = \pi(x_i) = P(Y_i = 1|X_i = x_i) = \frac{\exp(x_i^T \tilde{\beta})}{1 + \exp(x_i^T \tilde{\beta})}. \tag{C.2}$$

A probabilidade de falha é dada por $1 - \pi_i$, e o logaritmo da função de verossimilhança é dado por

$$\mathcal{L}(\tilde{\beta}) = \sum_{i=1}^{n} y_i x_i^T \tilde{\beta} - \ln(1 - x_i^T \tilde{\beta}). \tag{C.3}$$

O valor que maximiza $\mathcal{L}(\tilde{\beta})$ é obtido pela derivação $\mathcal{L}(\tilde{\beta})$ em relação a cada um de seus parâmetros. A solução do problema de otimização resultante não é uma expressão fechada. É necessário que seja realizado um processo iterativo para alcançar os valores de $\tilde{\beta}$ que maximizem a probabilidade. Nos resultados experimentais relatados neste artigo, consideramos o processo iterativo padrão usado por `scikit-learn` [13], conforme implementado em `liblinear` [22].

C.1.2 Máquinas de Vetores de Suporte

O classificador Máquinas de Vetores de Suporte (*Support Vector Machines* - SVM) foi proposto pela primeira vez para resolver problemas de reconhecimento de padrões. Os dados são mapeados em um espaço de maior dimensão e um hiperplano é construído de modo a separar os pontos em duas classes [53]. Os vetores que

definem a distância máxima entre as classes são chamados de *vetores de suporte*. O objetivo dos algoritmos relacionados ao SVM é encontrar o hiperplano que maximiza a separação entre as classes.

No caso linearmente separável, a ideia chave por trás de um SVM é bastante simples. Dado um conjunto de treinamento S contendo pontos das duas classes, o SVM os separa por um hiperplano, determinado por um certo subconjunto de pontos de S (os vetores de suporte). O hiperplano que separa os pontos possui a mesma distância entre as duas classes. Todos os vetores de suporte devem cumprir a margem mínima, sendo chamados vetores de margem. É comum encontrar situações onde as classes não são linearmente separáveis. Quando ocorre essa situação, os vetores do hiperplano e suporte são obtidos como solução para um problema de otimização com restrições [44].

Para separar conjuntos de dados separáveis linear e não linearmente, um ingrediente chave para os métodos SVM é uma função *kernel*. Usando uma função de *kernel*, o SVM projeta os dados em uma dimensão superior, onde se torna possível encontrar um hiperplano que separe as classes. Algumas das seguintes funções de *kernel* foram usadas pelos métodos SVM [28]:

- linear:

$$k(x_i, x_j) = x_i^T x_j \tag{C.4}$$

- função de base radial (RBF):

$$k(x_i, x_j) = \exp\left(-\frac{1}{2\gamma^2} \parallel x_i - x_j \parallel^2\right), \gamma > 0 \tag{C.5}$$

- polinomial:

$$k(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \tag{C.6}$$

- sigmóide:

$$k(x_i, x_j) = \tanh(\gamma x_i^T x_j + r). \tag{C.7}$$

Pequenos valores de γ tornam a superfície quase linear, enquanto aumenta seu valor aumenta a flexibilidade da superfície de decisão. O parâmetro r controla o limiar de deslocamento nos casos de núcleos polinomiais e sigmoides. O parâmetro d dá o grau do *kernel* polinomial. Todos os parâmetros devem ser especificados a priori pelo usuário.

C.1.3 *Naive Bayes*

O classificador *Naive Bayes* é amplamente utilizado na literatura como classificador de textos [37]. Utilizando probabilidade condicional entre as palavras do documento, pode ser encontrada a probabilidade de um documento pertencer à uma classe. A classe com maior probabilidade é escolhida para o documento [33].

Um classificador Bayesiano leva em consideração que a probabilidade da posição de um termo no documento é independente de sua posição verdadeira. O classificador considera que a probabilidade de um conjunto de palavras é o produtório entre as probabilidades das palavras. Desta forma, a formulação básica na técnica *Naive Bayes* é dada por:

$$V_{NB} = \text{argmax}_{v_j \epsilon V} \prod_{a_i \in \mathcal{D}} P\left(a_i | v_j\right) P(v_j) \tag{C.8}$$

onde V_{NB} é a categoria estimada, V é o conjunto de categorias, v_j é uma das categorias que pertence a V, a_i é uma palavra no documento $\mathcal{D}$, $P(v_j)$ é a

probabilidade *a priori* de v_j e $P(a_i|v_j)$ é a probabilidade de ocorrência de a_i na categoria v_j.

C.1.4 *Gradient Tree Boosting*

Utiliza um conjunto de árvores de decisão combinadas para aumentar o poder de classificação, através de um comitê. É utilizada uma abordagem sequencial, onde os erros de um passo são utilizados para reduzir possíveis ruídos para o próximo passo. O algoritmo *Gradient Tree Boosting* (GB) busca reduzir o erro das previsões utilizando a média do conjunto de treinamento inicial $f_0(x) = \bar{y}$ como solução e, iterativamente, atualiza modelo com o novo valor estimado multiplicado por um coeficiente de aprendizado, fazendo com que o modelo siga para direção inversa do gradiente da função objetivo $\Psi(y_i, f(x_i))$, até que uma condição de parada seja estipulada [36].

C.2 MÉTRICAS DE AVALIAÇÃO

A avaliação é muito importante para o desenvolvimento de um método de classificação de texto. Uma avaliação adequada, juntamente com uma análise comparativa, proporciona a informação de o quanto um classificador de texto proposto é bom. A fase de avaliação é a última e é fundamental para validar um método de classificação.

A tabela C.1 apresenta um matriz de confusão, que facilita o entendimento sobre o relacionamento entre a classe verdadeira de um documento e o resultado da resposta do classificador. Os possíveis relacionamentos são denominados a seguir:

Tabela C.1: Matriz de confusão

Classificação Real	Positivo	Negativo
Classificado como Positivo	Verdadeiro-Positivo	Falso-Positivo
Classificado como Negativo	Falso-Negativo	Verdadeiro-Negativo

- Verdadeiro-positivo (VP): Quantitativo de documentos que foram classificados como positivos corretamente.
- Falso-positivo (FP): Quantitativo de documentos que foram classificados como positivos incorretamente.
- Falso-negativo (FN): Quantitativo de documentos que foram classificados como negativos incorretamente.
- Verdadeiro-negativo (VN): Quantitativo de documentos que foram classificados como negativos corretamente.

Os relacionamentos descritos serão utilizados nas definições posteriores.

C.2.1 Acurácia e Erro

A acurácia (A) define o percentual de quantos documentos foram classificados corretamente pelo classificador dentre o total de documentos avaliados. O erro (E) representa os documentos classificados incorretamente dentre todos avaliados.

$$A = \frac{VP + VN}{VP + FP + FN + VN} \tag{C.9}$$

$$E = \frac{FP + FN}{VP + FP + FN + VN} \tag{C.10}$$

É importante ressaltar que:

$$Acurcia + Erro = 1 \tag{C.11}$$

C.2.2 Precisão e Revocação

Precisão e Revocação são medidas utilizadas para representar a qualidade do classificador. A precisão representa o percentual de documentos classificados corretamente como positivo sobre o total de documentos classificados como positivo.

$$P = \frac{VP}{VP + FP} \tag{C.12}$$

A revocação (R), que também é denominada por sensibilidade do classificador, representa o percentual de documentos classificados como positivo dentre todos os documentos que realmente são positivos.

$$R = \frac{VP}{VP + FN} \tag{C.13}$$

C.2.3 Validação Cruzada

De acordo com [39], a validação cruzada é um método considerado como padrão para validação estatística dos resultados. Ela consiste na divisão do conjunto de dados em k partes iguais ou bem próximas. Posteriormente, k iterações são realizadas, onde para cada iteração, uma parte dos dados são utilizados para teste

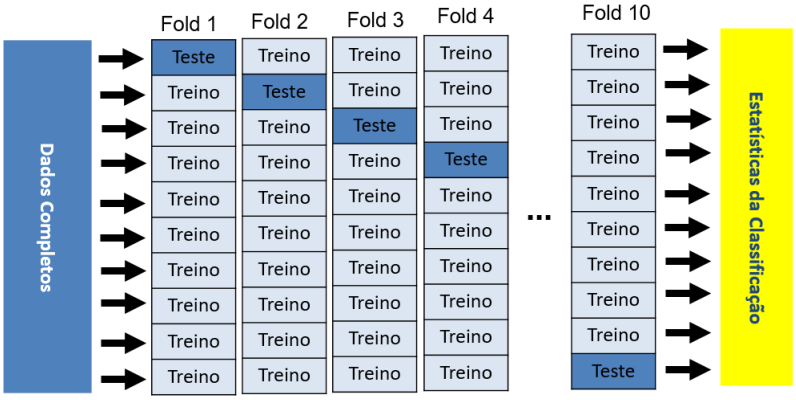

Figura C.1: Método *k-fold cross validation*

e $k-1$ são utilizados para o treinamento [48]. Ao final das iterações, a média das medidas de cada iteração é avaliada. O valor mais utilizado na literatura de k é 10, e, em tal situação, o método é chamado de validação cruzada do tipo *ten-fold* [6]. Nos casos onde os dados estão desbalanceados, é interessante particioná-los proporcionalmente.

Na figura C.1, pode ser observado o funcionamento do método, com os dados completos sendo divididos em 10 partes e para cada iteração, uma parte sendo utilizada para teste e as restantes para o para o treinamento. Ao final, a média das classificações de cada iteração é avaliada.