UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Vanderson Santana de Oliveira Leite Sampaio

5-Ions - Um Método para Estimar o Desempenho de Entidades a partir de Menções a Entidades Relacionadas em Textos na Web

Florianópolis
2019

Vanderson Santana de Oliveira Leite Sampaio

5-Ions - Um Método para Estimar o Desempenho de Entidades a partir de Menções a Entidades Relacionadas em Textos na Web

Dissertação submetida ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Ciências da Computação.
Orientador: Prof. Renato Fileto, Dr.
Coorientador: Prof. Douglas Dyllon Jeronimo de Macedo, Dr.

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Sampaio, Vanderson Santana de Oliveira Leite
5-Ions - Um método para estimar o desempenho de entidades a partir de menções a entidades relacionadas em textos na Web / Vanderson Santana de Oliveira Leite Sampaio ; orientador, Renato Fileto, coorientador, Douglas Dyllon Jeronimo de Macedo, 2019.
87 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2019.

Inclui referências.

1. Ciência da Computação. 2. Correlação de desempenho de entidades. 3. Predição de desempenho de entidades. 4. Anotações semânticas. 5. Relações semânticas. I. Fileto, Renato. II. Macedo, Douglas Dyllon Jeronimo de. III. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Ciência da Computação. IV. Título.

Vanderson Santana de Oliveira Leite Sampaio

5-Ions - Um Método para Estimar o Desempenho de Entidades a partir de Menções a Entidades Relacionadas em Textos na Web

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof.(a) Karin Becker, Dr(a).
Universidade Federal do Rio Grande do Sul

Prof. José Leomar Todesco, Dr.
Universidade Federal de Santa Catarina

Prof. Mauro Roisenberg, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciências da Computação.

Documento assinado digitalmente
Antonio Augusto Medeiros Frohlich
Data: 17/01/2020 12:45:50-0300
CPF: 500.696.400-68

Prof. Antônio Augusto Medeiros Fröhlich, Dr.
Coordernação do Programa

Documento assinado digitalmente
Renato Fileto
Data: 17/01/2020 11:29:54-0300
CPF: 106.315.998-94

Prof. Renato Fileto, Dr.
Orientador

Florianópolis, 19 de Novembro de 2019.

Este trabalho é dedicado a uma nação que crer em tempos melhores ignorando as mazelas de outrora.

AGRADECIMENTOS

Ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Santa Catarina e aos professores que nele estão e me acompanharam nessa jornada. Em especial aos professores doutores Renato Fileto e Douglas Dyllon Jeronimo de Macedo, pela paciência e orientação no desenvolvimento desta pesquisa. A Renato Fileto pelos ensinamentos, acadêmico e sociais, que permitiram expandir minha visão de mundo, sempre de maneira organizada, comprometida, responsável e antes de tudo de maneira correta e ética. A Douglas Macedo pelo apoio emocional e motivacional antes de tudo, que foi fundamental nessa jornada. Estendo ainda esse agradecimento a sua esposa e filha, Veridiana e Cecília, pessoas pela qual nutro uma relação de afeto familiar. Obrigado professores por me “adotarem”.

Agradeço a minha mãe, Veralucia Santana de Oliveira Sampaio, minha fonte de inspiração diária, por quem muito chorei e fiz chorar devido a distância física que nos separa. Dedico essa conquista à senhora. Eu jamais conseguiria sem você. Ao meu pai, Aloisio Leite Sampaio, por me ensinar o real valor das coisas. Aos meus irmãos Aloisio Leite Sampaio Júnior e Flávia Luciana Santana Leite Sampaio, com os quais compartilho meus momentos de alegrias e tristezas. A Flávia Luciana agradeço ainda por todas as conversas e apoios incondicionais, por trazer alegrias em momentos escuros, obrigado por ser quem você é. Aos meus irmãos mais distantes, porém não menos importantes, Vitória Leite Sampaio, Fabio Leite Sampaio, Valéria Leite Sampaio, Alisson Leite Sampaio e Flávio Leite Sampaio (in memoriam). Agradeço ainda a minha prima Rosaline por me presentar com meu pequeno e querido primo Davi Santana Costa, quem me proporciona momentos de alegria e descontração, suas mensagens, fotos e vídeos confortam meu espírito. Por fim, aos meus cachorros, Tobby (in memoriam) e Bapho, minhas gatas, Preta (in memoriam) e Veia, e ao papagaio Louro (in memoriam).

A Mariana Ribeiro, amiga que tive o prazer de reencontrar em Florianópolis, que sempre escuta minhas reclamações, histórias e brincadeiras, e por vezes exercita sua paciência mas com quem eu posso contar.

Em especial ao meu grande amigo Italo Otávio por ter aceitado o desafio de sair da cidade natal para me acompanhar por um tempo nessa trajetória. Não teria conseguido sem sua presença e apoio. Agradeço ainda a sua mãe Lenice Alves (in memoriam) por ter permitido que seu filho me acompanhasse no começo dessa aventura.

A todos os meus amigos, que a vida me deu e com quem sempre poderei contar, independente do momento e distância. A Suyaluane Italla, Dimas, Raphael, Cleidson, Evaldo, Janderson, Sidney, Ivo, Matheus Mendonça, Matheus Costa, Maicon Matheus, Ytallo, Marcos Emanuel, Marcos Aurélio, Allan, Keyzer (Baby), Renata Varjão e Tatyane.

Obrigado a todos os outros não mencionados mas que engradecem minha vida.

Aos meus colegas da Federação das Indústrias de Santa Catarina, em especial a Juliano Anderson Pacheco por ter apostado e lutado por minha causa. Um grande amigo que encontrei e levo para a vida. E a Dorzeli pela confiança e apoio no momento que mais precisei. Até ao próximo desafio.

"Behind every beautiful face there's
been some kind of pain."
(Bob Dylan)

RESUMO

Publicações na Web (e.g. notícias) podem influenciar a opinião pública acerca de certas entidades (e.g., políticos, instituições). Vários indicadores podem ser automaticamente extraídos dos textos dessas publicações e usados para estimar o comportamento do desempenho das entidades (e.g., popularidade, intenção de votos) ao longo do tempo. Este trabalho propõe um método automático que utiliza ferramentas do estado da arte em processamento de linguagem natural para identificar menções a entidades em textos e os sentimentos a elas associados. A partir dessas informações o método proposto calcula métricas que são usadas para construir modelos de regressão e de classificação para estimar tendências de desempenho das entidades mencionadas ou de entidades semanticamente relacionadas a elas. Nosso método calcula métricas de desempenho a partir de indicadores consolidados para entidades semanticamente relacionadas, avalia as correlações dessas métricas consolidadas com o desempenho real das entidades e usa essas métricas consolidadas para estimar o comportamento do desempenho de cada entidade. Um algoritmo genético, alguns métodos de classificação e técnicas de regressão foram usados para compor tais métricas consolidadas e efetuar predições de maneiras adequadas. Resultados experimentais em estudos de caso envolvendo política e economia mostram que métricas consolidadas para várias entidades inter-relacionadas são melhor correlacionadas com medidas reais de desempenho observadas para algumas entidades-alvo e levam a melhores previsões, em comparação com métricas para apenas uma entidade.

Palavras-chave: Correlação de desempenho de entidades. Predição de desempenho de entidades. Anotações semânticas. Relações semânticas.

ABSTRACT

Publications on the Web (e.g. news) may influence public opinion about certain entities (e.g., politicians, institutions). Various indicators can be automatically extracted from the texts of these publications and used to estimate entity performance (e.g., popularity, vote intention) over time. This paper proposes an automatic method that uses state-of-the-art tools for natural language processing to identify references to entities in texts and the associated sentiment. The extracted information is used to calculate metrics that are used to build regression and classification models to estimate the performance trends of the mentioned entities or entities semantically related to them. Our method calculates performance metrics from consolidated indicators for semantically related entities, assesses the correlations of these consolidated metrics with actual entity performance, and uses the consolidated metrics to estimate the performance of each entity. A genetic algorithm, some classification methods, and regression techniques were used to compose such consolidated metrics and make predictions in appropriate ways. Experimental results in case studies involving politics and economics show that consolidated metrics for several interrelated entities are better correlated with actual performance measures observed for some target entities and lead to better prediction than metrics for just one entity.

Keywords: Entity performance correlation. Entity performance prediction. Semantic Annotations. Semantic relatedness.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AfD Alternative für Deutschland

AG Algoritmo Genético

BACEN Banco Central

BC Base de Conhecimento

BI Business Intelligence

CDU Christlich-Demokratische Union Deutschlands

CNL Cloud Natural Language

CSU Christlich-Soziale Union in Bayern

DW Deutsche Welle

e.g. exempli grata

FDP Freie Demokratische Partei

GC Grafo de Conhecimento

Google CNL Google Cloud Natural Language

Google KG Google Knowledge Graph

GRUNE Bündnis 90Die Grünen

i.e. id est

IDE Integrated Development Environment

LINKE Die Linke

LSA Latent Semantic Analysis

NED Named Entity Desambiguation

NER Named Entity Recognition

OneR One Rule

PLN Processamento de Linguagem Natural

RLM Regressão Linear Múltipla

RLS Regressão Linear Simples

RS Rede Semântica

SPD Sozialdemokratische Partei Deutschlands

SVM Support Vector Machine

SUMÁRIO

1 INTRODUÇÃO

Publicações online (e.g., jornais, revistas), comentários associados a essas publicações e em mídias sociais, entre outras fontes de dados textuais na Web, podem ser valiosas para estimar alguma medida real de desempenho (e.g., intenção de voto em algum partido ou político, cotação da moeda, valor de mercado de uma marca ou empresa) associada as entidades mencionadas em tais textos. Algumas dessas entidades podem ter suas medidas reais de desempenho monitoradas, ao menos durante algum período de tempo, tornando possível avaliar a eficácia de estimativas dessas medidas a partir de fontes de dados online. Certas entidades são bastante mencionadas nessas fontes de dados em certos períodos de tempo, enquanto outras não. Portanto, um método para estimar a medida real de desempenho de uma entidade a partir de menções a ela e a entidades semanticamente relacionadas (i.e. entidades que possuem relações semânticas entre si, por exemplo, partidos políticos e seus membros filiados) pode ser útil para obter boas estimativas de medidas reais de desempenho, independente da pouca quantidade de menções, além de permitir avaliar a qualidade das estimativas, pelo menos para entidades cuja as medidas reais de desempenho estão disponíveis.

A Tabela 1 apresenta algumas características extraídas da publicação online "Os opositores de Merkel na União encontraram organização guarda-chuva"[1] (título original: *Merkel-Gegner in der Union gründen Dachverband*). O texto foi publicado pela Deutsche Welle (DW) no dia 25 de março de 2017. A tabela contém três colunas, a primeira representando as características avaliadas, seguido pelos valores obtidos ao se analisar o documento integralmente (Original). Por fim, a terceira coluna, Sumarização SumBasic, apresenta os resultados obtidos para o texto após ser submetido ao algoritmo de sumarização SumBasic (Capítulo 2), com compressão de 30%.

Inicialmente o texto original, escrito em Alemão, foi traduzido para o Inglês e então submetido a ferramenta de anotação Google Cloud Natural Language[2]. Essa tradução ocorre devido a limitações da ferramenta de anotação utilizada. As primeiras linhas da tabela demonstram a diminuição do tamanho do texto original para o texto sumarizado. Porém mesmo com a diminuição do texto algumas das principais características do texto original são mantidas em seu sumário, por exemplo, as principais entidades anotadas, o sentimento geral do documento e o sentimento associado a entidades Angela Merkel.

Várias características podem ser extraídas de textos (e.g., menções a entidades, sentimentos associados a elas ou aos trechos de texto onde ocorrem) para calcular indicadores de desempenho (e.g. quantidade de menções a uma entidade indicam a popularidade da entidade) e estimar alguma medida indireta (no escopo deste trabalho

[1] http://www.dw.com/de/merkel-gegner-in-der-union-gr%C3%BCnden-dachverband/a-38121446
[2] https://cloud.google.com/natural-language/

Tabela 1 – Características de uma publicação do Deutsche Welle.

Características	Original	Sumarização SumBasic
Parágrafos	6	1
Sentenças	20	6
Linhas	30	6
Palavras	387	86
Caracteres com Espaço	2.410	514
Caracteres sem Espaço	2.029	429
Quantidade de Entidades Anotadas	14	5
Algumas Entidades Presentes	CDU, Angela Merkel, CSU, Mitsch	CDU, Angela Merkel, Mitsch
Sentimento do Documento	Neutro	Neutro
Sentimento de Angela Merkel	Neutro	Neutro

Fonte: elaborado pelo autor.

essas medidas indiretas são chamadas de métricas) de desempenho da entidade (i.e., medidas calculadas a partir de indicadores que estão associadas as entidades), que pode ser comparada a medidas reais de desempenho, através da análise de sua correlação, por exemplo. Independentemente da existência de correlação entre métricas de desempenho calculadas e medidas reais de desempenho de uma entidade pode ser possível prever sua medida real desempenho. Os trabalhos (AHMED; SKORIC, 2014; FINK *et al.*, 2013) investigam correlações entre indicadores de desempenho extraídos de *tweets* e medidas reais de desempenho de algumas entidades. Enquanto o método descrito em (AHMED; SKORIC, 2014) depende apenas da popularidade (quantidade de menções) de entidades no texto, (FINK *et al.*, 2013) também considera a polaridade dos sentimentos (i.e. menções com sentimentos positivos ou negativos) para calcular as métricas de desempenho. Outros trabalhos, como (TUMITAN; BECKER, 2014; RAMTEKE *et al.*, 2016), preveem variações de desempenho para determinadas entidades. O método proposto em (RAMTEKE *et al.*, 2016) foi capaz de prever o vencedor de uma eleição, analisando a popularidade e a polaridade dos sentimentos das menções aos candidatos. Enquanto isso, (TUMITAN; BECKER, 2014) aplicou aprendizado de máquina para prever variações na intenção de voto de determinados partidos e políticos.

Entretanto, esses trabalhos usam apenas dicionários de nomes de superfície, i.e., nomes que representam alguma entidade (e.g. Dilma, Dilmãe, Dilminha são alguns nomes de superfície para a entidade Dilma Rousseff), para identificar entidades no texto. Desta maneira, eles não aproveitam os benefícios das ferramentas de anotação semântica do estado da arte, como: maior precisão e cobertura, identificação de correferências e exploração de conexões entre entidades em redes semânticas. Além disso,

eles não exploram possíveis influências de entidades semanticamente relacionadas entre si, por exemplo, agregando seus indicadores de desempenho para calcular as métricas de desempenho consolidadas. Em outras palavras, eles não investigam como indicadores e métricas de desempenho, que combinam menções e seus sentimentos para conjuntos de entidades semanticamente relacionadas (e.g. políticos do mesmo partido), podem estar associados ao desempenho de entidades de tais conjuntos, por exemplo, através do cálculo de correlação ou de predição.

Desta maneira, este trabalho busca elucidar as seguintes perguntas de pesquisa:

1. O uso de ferramentas do estado da arte em Processamento de Linguagem Natural (PLN) para identificar, selecionar e desambiguar anotações semânticas pode contribuir para melhorar correlações e predições de desempenho de entidades usando métricas calculadas a partir de tais menções e sentimentos extraídos de textos?
2. Quais os impactos da sumarização de texto na análise de correlação entre as métricas calculadas a partir das características extraídas de textos e as medidas reais de desempenho de uma entidade? Sumarização pode contribuir para produzir melhores correlações e predições e/ou reduzir o tempo total de processamento para obtê-las?
3. Menções a entidades relacionadas a uma entidade alvo podem contribuir para a correlação com medidas reais desempenho e para a predição de desempenho da entidade alvo?

1.1 OBJETIVOS

O objetivo deste trabalho é desenvolver um método que realiza extração de informação de textos para calcular métricas de desempenho para um conjunto de entidades relacionadas a alguma entidade alvo. Tais métricas são utilizadas na correlação e predição de medidas reais de desempenho da entidade alvo.

1.1.1 Objetivos Específicos

1. Entender e aplicar métodos e ferramentas do estado da arte em sumarização de texto, ligação de entidades (incluindo correferências) e análise de sentimentos para calcular métricas de popularidade e sentimento associado a menções as entidades em textos.
2. Estudar, selecionar e adaptar métricas calculadas a partir de tais características extraídas dos textos, inclusive, consolidando as métricas para conjuntos de entidades semanticamente relacionadas.

3. Estudar, selecionar e desenvolver técnicas, tais como algoritmos genéticos, para estimar a relevância (peso) das entidades semanticamente relacionadas a uma entidade alvo no cálculo de métricas consolidadas.

4. Correlacionar as métricas consolidadas com as medidas reais de desempenho de entidades alvo.

5. Desenvolver um método para predizer o desempenho de entidades com base nas métricas consolidadas para conjuntos de entidades semanticamente relacionadas.

6. Aplicar o método proposto em estudos de caso para domínios de aplicação tais como política e negócios.

1.2 METODOLOGIA

Este trabalho envolve uma pesquisa aplicada, cuja motivação é a necessidade de produzir uma solução na forma de artefato de software para análise de dados, além de avaliar o seu desempenho. A natureza aplicada desta pesquisa não elimina a necessidade de uma fundamentação teórica, que é necessária para prospectar possibilidades de solução do problema abordado e posteriormente auxiliar na interpretação e na análise dos resultados de experimentos. Assim, o trabalho foi proposto a partir de uma fundamentação teórica que envolveu busca bibliográfica sistemática por trabalhos com temas correlatos ao desta pesquisa. A fundamentação abordou também técnicas e ferramentas para extração de informação de texto, utilização de redes semânticas para determinar conjuntos de entidades semanticamente relacionadas, análise de correlação e análise preditiva do desempenho de entidades. Daí foram desenvolvidos uma proposta de solução e um plano de experimentos para avaliar as questões de pesquisa levantadas na fundamentação teórica. Uma ferramenta foi implementada para executar o método proposto, permitindo realizar experimentos visando a avaliação do desempenho da proposta. As métricas de desempenho consolidadas para entidades semanticamente relacionadas e as predições derivadas de tais métricas consolidadas foram comparadas com medidas reais de desempenho de certas entidades, para avaliar correlações entre métricas calculadas e reais e a acurácia das predições realizadas usando tais métricas, de modo a responder nossas perguntas de pesquisa.

1.3 CONTRIBUIÇÕES E LIMITAÇÕES

As principais contribuições deste trabalho são: (i) utilização de redes semânticas para determinar entidades relacionadas e calcular métricas que consideram todas as entidades relacionadas a uma entidade alvo para estimar alguma medida real

de desempenho, independentemente do número de menções a entidade alvo; (ii) a utilização de algoritmos de busca por soluções e otimização (e.g. algoritmo genético), como uma alternativa para ponderar a influência das entidades relacionadas a alguma entidade alvo, essa solução é útil, principalmente, para classificação de tendências de classes de valores categóricos (e.g, aumenta, diminui, mantém); (iii) uso de técnicas de regressão para ponderar a influência das entidades relacionadas quando é realizado a predição de variáveis contínuas de desempenho; e (iv) o fluxo de atividades do método *5-Ions* proposto, que permite extrair características de textos usando ferramentas do estado da arte, além de calcular métricas de desempenho, levando em consideração entidades semanticamente relacionadas, para analisar sua correlação com medidas reais desempenho de entidades e predizer tais medidas de desempenho.

Experimentos em estudos de casos envolvendo popularidade em política e cotação da moeda brasileira mostraram bons resultados ao se analisar o desempenho de uma entidade alvo considerando menções a ela e a entidade relacionadas em uma rede semântica. As métricas de desempenho consolidadas sempre se mostraram superiores às métricas restritas às menções e aos sentimentos da entidade alvo, tanto em termos de correlação com medidas reais de desempenho quanto qualidade das predições. Já a utilização de sumarização de texto propiciou ganhos no tempo de processamento, porém sem garantir melhores resultados de correlação e predição.

As limitações desse trabalho incluem: (i) não houve implementação de nenhuma técnica para extração de informação de textos nem para ligação de entidades (*Entity Linking - EL*), sendo usadas ferramentas existentes para realizar tais tarefas; (ii) métodos alternativos para sumarização de texto foram pouco explorados e somente em um estudo de caso, havendo potencial para ganhos adicionais; (iii) o trabalho limitou-se a aplicar regressão linear simples, deixando para trabalhos futuros a possibilidade de aplicar técnicas como regressão linear múltipla, regressão de Rigde e regressão Lasso; e (iv) faltam mais comparações com trabalhos relacionados a fim de melhor avaliar o uso de ferramentas do estado da arte para extração de características de texto e EL pelo nosso método, frente a métodos que utilizam somente identificação de entidades baseada em dicionários, por exemplo.

1.4 ESTRUTURA DO TRABALHO

O restante deste trabalho está organizado da seguinte maneira. O Capítulo 2 aborda a fundamentação teórica necessária para a compreensão deste trabalho. O Capítulo 3 relaciona trabalhos correlatos recentes ao que é desenvolvido nesta pesquisa. O Capítulo 4 detalha nossa proposta de pesquisa, com a apresentação e a descrição do processo proposto para o cálculo de métricas visando investigar correlações e predição de desempenho. O Capítulo 5 detalha o plano de experimentos, bem como as configurações utilizadas. O Capítulo 6 exibe e discute os resultados obtidos

nos experimentos. Por fim, o Capítulo 7 apresenta as conclusões desta pesquisa e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica necessária para o entendimento deste trabalho. Este está dividido em seis seções. A Seção 2.1 aborda as redes semânticas, incluindo algumas técnicas para extração de redes semânticas relevantes para análise de menções a entidades em texto. A Seção 2.2 apresenta um panorama de tarefas, técnicas e ferramentas para extração de informação em textos, com explanações sobre reconhecimento e desambiguação de entidades nomeadas, análise e classificação de sentimentos e sumarização automática de texto. A Seção 2.3 introduz o conceito de indicadores e métricas de desempenho, o qual é aprofundado no decorrer deste documento. A Seção 2.4 apresenta algoritmos de otimização e busca para selecionar parâmetros em uma equação. A Seção 2.5 discorre sobre as análises de correlação e suas medidas. A Seção 2.6 trata de análises preditivas através de técnicas de regressão e de classificação. Por fim, a Seção 2.7 detalha quais as técnicas são utilizadas no decorrer do trabalho.

2.1 REDES SEMÂNTICAS

Uma entidade nomeada (NADEAU; SEKINE, 2007) é um objeto de alguma classe (e.g., pessoa, lugar, instituição, moeda) (VIEIRA; LIMA, 2001) e que tem certas propriedades de informação. Uma entidade pode ser mencionada em um texto através de algum dos seus nomes de superfície (RAO *et al.*, 2013) (e.g. “Merkel” e “Angela Merkel” são nomes de superfície da mesma entidade do tipo pessoa). Toda referência a uma entidade nomeada em texto é chamada de menção. O significado de uma menção é determinado pelo contexto no qual ela está inserida. O contexto textual de uma menção pode ser dado por palavras vizinhas, sentença ou parágrafo em que ela está situada, enquanto o seu contexto mais amplo refere-se a atributos como fonte de dados, local, período de tempo e autor.

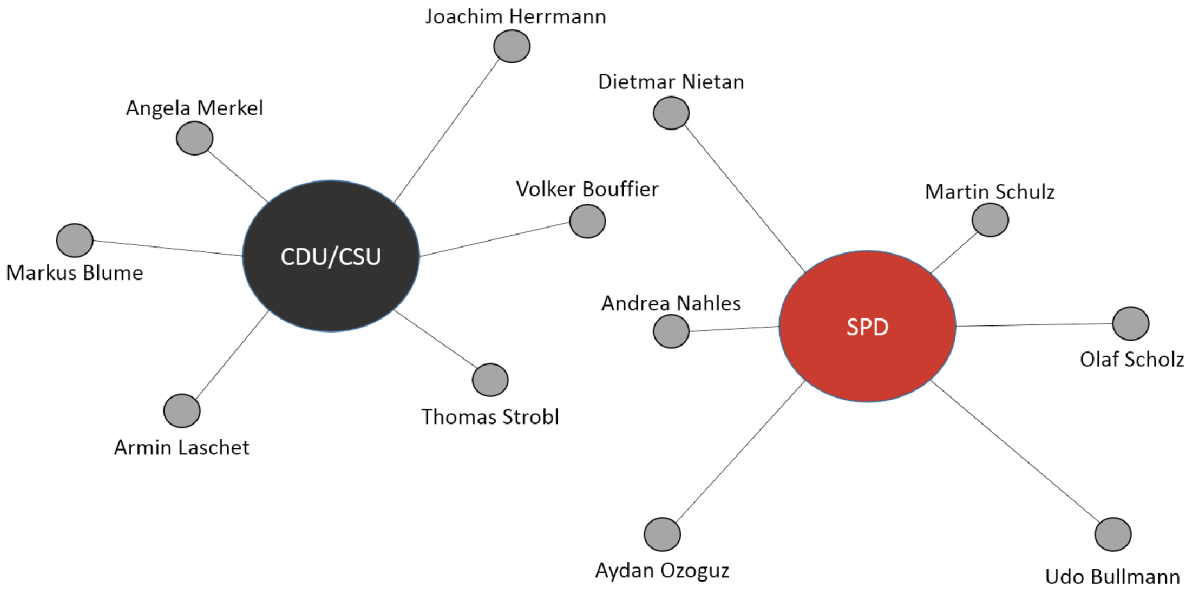
Entidades nomeadas podem ter várias relações entre si. Uma rede semântica (RS) (SOWA, 1991) é um grafo direcionado $RS(V, E)$ que representa o conhecimento como relações entre coisas. Cada vértice $v \in V$ da RS representa um recurso, tal como uma entidade nomeada, uma classe, um evento ou mesmo uma palavra simples ou composta com semântica bem definida. As arestas do conjunto $E \subseteq V \times V$ representam as relações semânticas entre esses recursos. Uma aresta direcionada $e \in E$ conecta um recurso $v \in V$ a algum outro recurso $v' \in V$ de acordo com alguma relação semântica. Exemplos de grandes RSs são os Grafos de Conhecimento (GC) como o Google Knowledge Graph[1] utilizado pela Google Cloud Natural Language[2], a

[1] https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html
[2] https://cloud.google.com/natural-language/

DBpedia[3] usado pela DBpedia-Spotlight (MENDES *et al.*, 2011) e a Babelnet[4] usado pela Babelfy (MORO *et al.*, 2014).

Neste trabalho, são usadas redes semânticas menores e com um número de componentes conexos, onde cada componente representa um grupo de entidades nomeadas semanticamente relacionadas. A direção das arestas é abstraída para determinar os componentes conexos. A Figura 1 mostra uma pequena RS extraída do GC DBpedia, com dois componentes conexos: um centralizado na união dos partidos CDU/CSU e outro centralizado no partido SPD. Os principais políticos Alemães são vinculados aos respectivos partidos (ou união de partidos) através da relação de filiação, representada por arestas da rede semântica ilustrada na Figura 1, com os respectivos componentes conexos centrados em CDU/CSU e SPD.

Figura 1 – Exemplo de rede semântica com componentes conexos referentes a partidos Alemães e seus principais filiados.



Fonte: elaborado pelo autor baseado na DBpedia.

2.1.1 Técnicas de construção de Redes Semânticas

Pode-se construir uma RS a partir de um grande GC, como a DBpedia, mediante a extração de entidades e relações semânticas relevantes entre essas entidades. Por exemplo, pode-se selecionar somente recursos (vértices) das classes partido político e seus políticos afiliados, sendo tal relação entre cada político e partido representada por uma aresta da relação afiliação.

[3] http://dbpedia.org
[4] https://babelnet.org/

Os componentes conexos de uma RS também podem ser obtidos de outras maneiras. Por exemplo, (RIBEIRO *et al.*, 2018) analisa escândalos de corrupção no Brasil e a partir deles constrói uma RS de pessoas e instituições envolvidas em cada caso de corrupção. O relacionamento entre as entidades, em tal RS surge quando duas entidades são citadas em um mesmo escândalo de corrupção. Em outras palavras, a RS pode ser gerada mediante a análise das ocorrências de menções a duas ou mais entidades em um mesmo conjunto de documentos. Outra alternativa para extrair componentes conexos de grandes GCs é agrupar seus recursos próximos (segundo alguma medida de distância) ou densamente conectados (MORO *et al.*, 2014).

2.2 EXTRAÇÃO DE INFORMAÇÃO DE TEXTO

Esta seção apresenta as técnicas de extração de informação em texto usadas no escopo deste trabalho: reconhecimento e desambiguação de entidades nomeadas, análise de sentimentos e sumarização automática de texto. As próximas subseções abordam cada uma dessas técnicas.

2.2.1 Reconhecimento e Desambiguação de Entidades Nomeadas

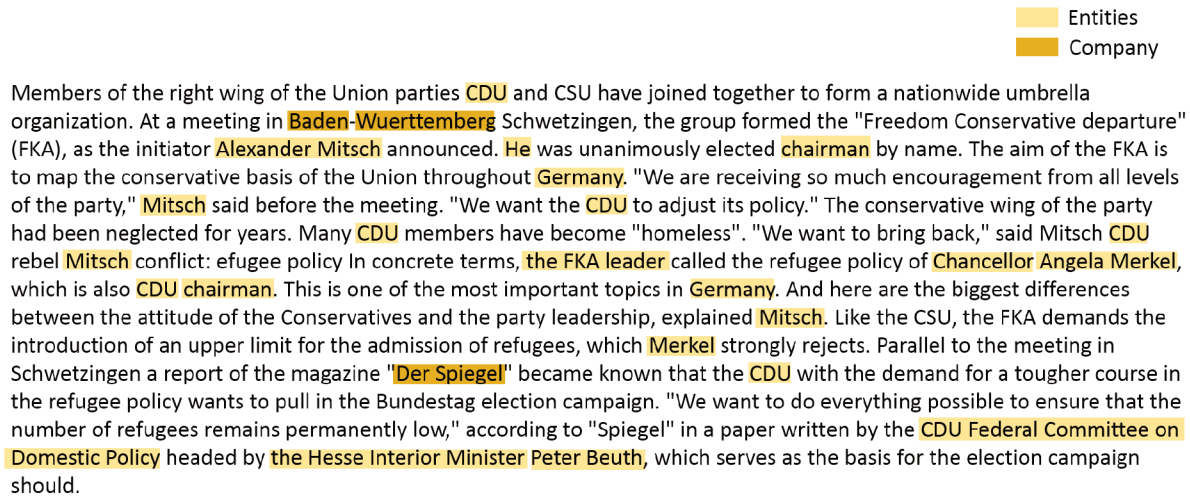
O Reconhecimento de Entidades Nomeadas (*Named Entity Recognition* - NER) é a tarefa de extração de informação que reconhece menções a entidades nomeadas em um texto (delimitando o início e o fim de cada menção no texto) e as classifica. Usualmente NER classifica as menções encontradas de acordo com um determinado conjunto de classes de entidades (NADEAU; SEKINE, 2007). A Desambiguação de Entidades Nomeadas (*Named Entity Desambiguation* - NED), também chamada de ligação de entidades (do inglês *Entity Linking* - *EL*), por sua vez, tenta vincular cada menção previamente reconhecida em um texto a sua correta definição em um banco de dados ou uma base de conhecimento (BC), de acordo com o contexto em que a menção está inserida (SHEN, Wei *et al.*, 2014). Atualmente, existe uma grande variedade de ferramentas e técnicas disponíveis para realizar NER/NED. Algumas delas ligam as menções reconhecidas em texto a recursos de uma BC (semi-)estruturada, como um GC ou uma rede semântica (MENDES *et al.*, 2011; SPECK; NGOMO, 2014; MORO *et al.*, 2014; SHEN, W. *et al.*, 2015).

A Figura 2 apresenta o resultado do reconhecimento de entidades nomeadas da ferramenta Open Calais[5]. Por se tratar de uma ferramenta de uso comercial, ela apresenta distinção entre as entidades do tipo Empresa e as demais entidades. As entidades do tipo Empresa estão destacadas, na figura, na cor marrom e as demais entidades estão destacadas na cor bege. A Figura 2 também permite identificar limitações de cobertura (i.e. capacidade de anotar todas as entidades nomeadas presentes no

[5] http://www.opencalais.com/opencalais-demo/

texto) da técnica de anotação empregada pelo Open Calais. Por exemplo, a entidade CSU presente no texto não é anotada por essa ferramenta.

Figura 2 – Exemplo de anotação textual efetuada pela ferramenta Open Calais.



Fonte: elaborado pelo autor baseado no Open Calais.

A ocorrência de entidades nomeadas em texto pode envolver alguns fenômenos, tais como: polissemia, sinonímia e correferência. A polissemia ocorre quando algum nome de superfície pode se referir a diferentes coisas (e.g. “Merkel” pode se referir à chanceler Alemã ou a uma cidade no Texas). Sinonímia ocorre quando nomes de superfícies diferentes se referem à mesma coisa. Correferências, por outro lado, são menções a uma mesma entidade representadas por palavras distintas que não são nomes de superfície da referida entidade (PRADHAN *et al.*, 2011). Por exemplo, na Figura 2 o pronome pessoal "*He*"(ele), que inicia uma frase na terceira linha do texto, refere-se a Alexander Mitsch.

2.2.2 Análise de Sentimento

A análise de sentimento, também chamada de mineração de opinião, tem o objetivo de identificar a polaridade do sentimento expresso no texto. Usualmente, as soluções de análise de sentimento tentam classificar o sentimento em três categorias: positiva, negativa ou neutra. Essas soluções podem avaliar o sentimento em diferentes contextos textuais que podem cobrir: uma entidade; uma sentença ou frase; ou todo documento texto (LIU, 2012). Algumas ferramentas que aplicam técnicas de Processamento de Linguagem Natural (PLN) realizando tarefas como NER/NED (e.g., *Google Cloud Natural Language*[6], *IBM Watson Natural Language Understanding*[7]) também

[6] https://cloud.google.com/natural-language/?hl=pt-br
[7] https://www.ibm.com/watson/services/natural-language-understanding/

podem classificar o sentimento associado a alguma entidade específica ou a algum escopo textual maior (e.g., sentença, documento completo).

A análise de sentimento no contexto de entidade identifica as menções as entidades nomeadas contidas no texto e então define o sentimento associado a cada menção. O texto pode ter inúmeras menções a entidades com opiniões distintas (LIU, 2012). Já a análise de sentimento no contexto de sentença separa o texto de acordo com suas frases para então determinar o sentimento e a influência da sentença no contexto do texto completo (LIU, 2012). Cada sentença possui um sentimento. Por fim, a análise de sentimento no contexto de documento completo qualifica o sentimento geral contido em todo o documento texto analisado (LIU, 2012). O sentimento das sentenças, bem como suas influências, compõe o cálculo do sentimento do documento completo.

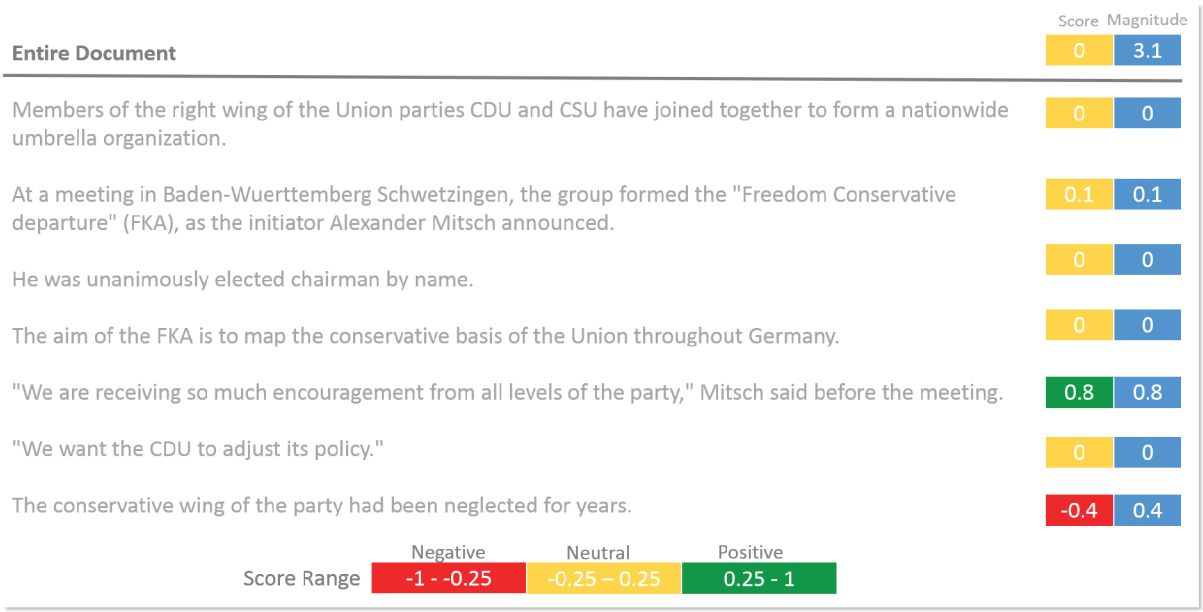
A Figura 3 ilustra o resultado da análise de sentimento da ferramenta *Google Cloud Natural Language* (Google CNL). A ferramenta classifica o sentimento como: negativo quando o valor pertence ao intervalo $[-1, -0.25[$; neutro quando o valor pertence ao intervalo $[-0.25, 0.25]$; e positivo quando o valor pertence ao intervalo $]0.25, 1]$. A Figura 3a exibe as análises de sentimentos nos contextos de documento e sentença. O documento possui um sentimento neutro (i.e. valor de sentimento igual a 0), com os sentimentos das sentenças variando entre positivo (e.g. valor de sentimento igual a 0.8), negativo (e.g. valor de sentimento igual a –0.4) e neutro. Já a Figura 3b representa a entidade nomeada Alexander Mitsch, objeto da classe Pessoa, com um sentimento neutro (valor igual a –0.1) associado a ela.

2.2.3 Sumarização Automática de Texto

A sumarização automática de texto é uma tarefa de PLN que gera resumos (sumários) de documentos de texto. Os resumos podem ter origem em um ou vários documentos, desde que preservem as informações importantes contidas nesses documentos (DAS; MARTINS, 2007). Há dois tipos de sumarização automática de texto: sumarização por extração e sumarização por abstração (DAS; MARTINS, 2007). A sumarização por abstração tenta compreender as sentenças contidas nos documentos textuais. Ela permite a substituição de frases, contanto que a nova frase permita um melhor entendimento. A complexidade da geração de linguagem natural dificulta a utilização da sumarização por abstração (DAS; MARTINS, 2007; SARANYAMOL; SINDHU, 2014).

Já a sumarização por extração simplesmente seleciona as principais sentenças dos documentos para gerar o sumário (SARANYAMOL; SINDHU, 2014). Dentre os vários métodos de sumarização por extração temos: Luhn, Sumbasic e *Latent Semantic Analysis* (LSA). O método Luhn utiliza a frequência das palavras e a distância entre elas para calcular a importância das frases que as contêm. O sumário é composto pelas

Figura 3 – Resultado da análise de sentimento da ferramenta Google CNL.



(a) Contexto de documento e de sentença

(b) Contexto de entidade

Fonte: elaborado pelo autor baseado no Google CNL.

frases mais importantes (LUHN, 1958). Já o método Sumbasic calcula a importância das frases de acordo com a frequência das palavras, apenas (VANDERWENDE *et al.*, 2007).

Por fim, o método LSA utiliza uma técnica algébrico-estatística. Essa técnica identifica as ocorrências de polissemia e sinonímia através dos valores estatísticos de coocorrência das palavras. A identificação desses fenômenos permite verificar a proximidade das frases independente das suas composições. Desta forma, as frases do sumário contêm relevância estatística e não apresentam redundância entre elas (DEERWESTER *et al.*, 1990).

2.3 INDICADORES E MÉTRICAS DE DESEMPENHO

No contexto deste trabalho, indicadores de desempenho são valores atômicos, de composição simples obtidos a partir da extração de alguma característica em texto em determinado período de tempo. As menções a entidades e os sentimentos associados a elas são algumas das características que podem ser extraídas dos textos.

Desta maneira, alguns indicadores de desempenho resultantes dessas características podem incluir: popularidade de uma entidade (número de menções) e frequência das polaridades dos sentimentos associados as menções às entidades, i.e. a frequência dos sentimentos positivo, negativo ou neutro associados as entidades (FINK *et al.*, 2013; AHMED; SKORIC, 2014).

Como visto neste capítulo, algumas técnicas e ferramentas de PLN calculam a intensidade, em valor numérico, do sentimento associado às menções. Isso permite o cálculo de métricas de desempenho, a partir da composição de indicadores. Métricas de desempenho mais complexas podem ser calculadas, por exemplo, através da razão das sumarizações dos sentimentos positivos e negativos (indicadores de polaridade) das entidades (TUMITAN; BECKER, 2014).

O Quadro 1 transcreve as métricas de desempenho definidas em (TUMITAN; BECKER, 2014). Nas fórmulas descritas no quadro: V representa o conjunto de todas as entidades em análise; v corresponde a um elemento do conjunto V; ρ é o período de tempo da análise; pos_V corresponde ao sentimento positivo associado as menções a entidade v; de maneira análoga neg_V representa o sentimento negativo associado as menções a entidade v.

Quadro 1 – Métricas de desempenho para entidades em período de tempo ρ.

Métrica	Fórmula	Descrição
f_1	$f_1(v,\rho) = \frac{\sum_{t\epsilon\rho} pos_{v,t}}{\sum_{t\epsilon\rho} neg_{v,t}}$	Razão do sentimento positivo em relação a uma entidade sobre o sentimento negativo da mesma entidade.
f_2	$f_2(v,\rho) = \frac{\sum_{t\epsilon\rho} pos_{v,t}}{\sum_{t\epsilon\rho} \langle pos_{v,t}+neg_{v,t}\rangle}$	Razão do sentimento positivo em relação a uma entidade sobre o sentimento total da mesma entidade.
f_3	$f_3(v,\rho) = \frac{\sum_{t\epsilon\rho} neg_{v,t}}{\sum_{t\epsilon\rho} \langle pos_{v,t}+neg_{v,t}\rangle}$	Razão do sentimento negativo em relação a uma entidade sobre o sentimento total da mesma entidade.
f_4	$f_4(v,\rho) = \frac{\sum_{t\epsilon\rho} \langle pos_{v,t}-neg_{v,t}\rangle}{\sum_{t\epsilon\rho} \langle pos_{v,t}+neg_{v,t}\rangle}$	Razão da diferença entre o sentimento positivo e negativo em relação a uma entidade sobre o sentimento total da mesma entidade.
f_5	$f_5(v,\rho) = \frac{\sum_{t\epsilon\rho} pos_{v,t}}{\sum_{q\epsilon V}\sum_{t\epsilon\rho} pos_{q,t}}$	Razão do sentimento positivo em relação a uma entidade sobre o sentimento positivo total (em relação a todas as entidades).
f_6	$f_6(v,\rho) = \frac{\sum_{t\epsilon\rho} neg_{v,t}}{\sum_{q\epsilon V}\sum_{t\epsilon\rho} neg_{q,t}}$	Razão do sentimento negativo em relação a uma entidade sobre o sentimento negativo total (em relação a todas as entidades).
f_7	$f_7(v,\rho) = \frac{\sum_{t\epsilon\rho} men_{v,t}}{\sum_{q\epsilon V}\sum_{t\epsilon\rho} men_{q,t}}$	Razão das menções em relação a uma entidade sobre o total de menções (em relação a todas as entidades).

Fonte: (TUMITAN; BECKER, 2014).

Ao todo são apresentados sete métricas onde seis dessas métricas combinam sentimentos positivos e negativos associados as entidades alvo e somente uma métrica considera a quantidade de menções as entidades. As quatro métricas iniciais analisam somente o sentimento associado as menções a uma entidade alvo, já as três métricas finais consideram o conjunto de métricas em análise (e.g. todos os candidatos de uma eleição) para então calcular o valor. Todas as métricas consideram a data de publicação dos textos no momento do cálculo. Perceba que todas as métricas consideram somente a entidade alvo, ignorando eventuais influências (positiva ou negativa) de outras entidades no seu cálculo.

2.4 ALGORITMOS DE BUSCA POR SOLUÇÕES E DE OTIMIZAÇÃO

Algoritmos de busca por soluções tentam encontrar soluções para problemas complexos. As buscas podem ser globais ou locais. Uma busca global necessita conhecer as etapas percorridas até encontrar uma solução, ou seja, todo o percurso (global) é importante para a solução. Já uma busca local, ignora as etapas percorridas, necessitando apenas atingir um objetivo válido, ou seja, apenas o estado atual (local) é importante para a solução (RUSSELL; NORVIG, 2003). Algoritmos de busca local podem ser utilizados para solucionar problemas de otimização, encontrando o melhor estado de acordo com uma função objetivo.

Encontrar variáveis (ou parâmetros) que maximizem alguma função pode demandar tempo e recursos computacionais elevados. Esses problemas podem ser solucionados através do uso de funções objetivo para orientar a busca no espaço de soluções (MITCHELL, 1998). Um Algoritmo Genético (AG) pode ser utilizado para encontrar soluções ótimas ou próximas de ótimas para os problemas de otimização de parâmetros. Para tal, os AGs utilizam conceitos de indivíduos, populações, transições estatísticas ou mutações e funções avaliação (LINDEN, 2008).

Os algoritmos genéticos são um ramo da computação evolucionária que modela e aplica o processo evolutivo dos seres vivos. Para tal, é necessário mapear os estados possíveis como indivíduos, que devem possuir uma representação cromossômica. Um conjunto de N indivíduos aleatórios formam uma população. Cada indivíduo possui um valor de aptidão (*fitness*) calculado a partir de uma função de avaliação (*fitness function*). A função avaliação busca mensurar a proximidade do indivíduo ao objetivo. Indivíduos com maior aptidão estão mais próximos de solucionar o problema apresentado. O valor de aptidão permite que os melhores indivíduos sejam selecionados para se reproduzirem. O processo de reprodução, ou cruzamento, gera novos indivíduos (filhos) a partir da seleção de dois indivíduos da população (pais). Os pais selecionados têm suas cargas cromossômicas combinadas a partir de um ponto de corte, resultado em dois filhos. A cada novo cruzamento existe uma probabilidade de ocorrer uma mutação, ou transição estatística, i.e. os novos indivíduos podem ter algum

dos seus cromossomos substituído de forma aleatória. Todos os filhos da população compõem uma nova geração da população. O número de gerações pode ser utilizado como critério de parada do algoritmo, de forma a garantir que a execução tenha um fim (LINDEN, 2008; RUSSELL; NORVIG, 2003).

2.5 ANÁLISE DE CORRELAÇÃO

A análise de correlação verifica a dependência (associação) entre variáveis (REVELLE, 2013). O coeficiente de correlação quantifica a associação mútua entre duas variáveis, i.e., quão forte é a relação entre os valores das variáveis. Os coeficientes são valores pertencentes ao intervalo $[-1, 1]$. Os extremos do intervalo (valores próximos a 1 ou -1) significam alta correlação (próximo a relações perfeitas). Por outro lado, valores próximos a zero indicam fraca correlação, ou até mesmo inexistência de correlação (REVELLE, 2013; BENESTY *et al.*, 2009). O sinal do coeficiente (positivo ou negativo) informa a direção da correlação. Assim, coeficientes maiores que zero (relação positiva) demonstram proporcionalidade direta entre as variáveis. De maneira análoga, coeficientes menores que zero (relação negativa) expõem proporcionalidade inversa entre as variáveis. Destaca-se que correlações expressam apenas a relação entre variáveis, e não uma relação de causa e efeito (BENESTY *et al.*, 2009).

Dentre as principais medidas de correlação temos: correlação de Pearson; correlação de Spearman; e correlação de Kendel. A correlação de Pearson é um coeficiente de correlação utilizado para variáveis quantitativas contínuas, com aplicações em técnicas de regressão (BENESTY *et al.*, 2009). Já as medidas de correlação de Spearman e Kendall são não lineares. A correlação de Spearman é aplicada a variáveis ordinais com distribuição desconhecida (GRZEGORZEWSKI; ZIEMBINSKA, 2011). Por fim, a correlação de Kendall é aplicada a um conjunto finito, e com tamanho não elevado, de amostras das variáveis (XU *et al.*, 2013).

2.6 ANÁLISE PREDITIVA

A análise preditiva examina dados atuais e históricos para realizar previsões. Tais dados são usados para criar e calibrar os modelos preditivos. Dentre os principais modelos de predição temos os modelos estatísticos, com as técnicas de regressão, e as técnicas baseadas em classificação de tendências.

2.6.1 Técnicas de Regressão

Técnicas de regressão possibilitam realizar análises preditivas de variáveis dependentes a partir de variáveis independentes. Por esse motivo, as técnicas de regressão são relações de causa e efeito entre as variáveis. Essas relações são expressas através de modelos matemáticos (YAN; SU, 2009). Regressões lineares são usadas

para encontrar relações lineares entre uma variável alvo e uma ou mais variáveis utilizadas para a sua predição. Quando a relação procurada da variável alvo é com apenas uma variável independente tem-se uma Regressão Linear Simples (RLS). Por outro lado, quando a relação procurada é com diversas variáveis independentes tem-se uma Regressão Linear Múltipla (RLM) (YAN; SU, 2009).

2.6.2 Técnicas de Classificação de Tendências

Algumas técnicas de aprendizado supervisionado utilizam dados históricos classificados para predizer as classes de novas ocorrências de dados. Tais dados podem ter valores discretos ou contínuos, enquanto suas classes referem-se somente a valores discretos (RUSSELL; NORVIG, 2003). Os algoritmos de aprendizado de máquina, naïve Bayes e OneR, descritos a seguir, são alguns dos mais utilizados e que apresentam bons resultados em classificação. Porém técnicas mais robustas, como *Support Vector Machine* (SVM), *Random Forest* e outros classificadores baseados em árvores, podem ser utilizados para classificação de tendências.

O algoritmo de classificação naïve Bayes é um classificador probabilístico supervisionado baseado no teorema de Bayes. O naïve Bayes, assim como o teorema de Bayes, considera as variáveis independentes no momento da predição (PATTEKARI; PARVEEN, 2012). Já o OneR (*One Rule*) é um simples, porém preciso, algoritmo de classificação supervisionado. O algoritmo gera uma regra para cada atributo de dado, para então selecionar a regra com o menor erro total. Daí o nome "One Rule"(uma regra). A classificação mais frequente para cada atributo é utilizada para criar a regra desse atributo (MAHAJAN; GANPATI, 2014).

2.7 CONSIDERAÇÕES FINAIS

O método *5-Ions*, proposto nesse trabalho (Capítulo 4), é constituído por algumas atividades distintas. Desta forma, essa seção destaca quais técnicas serão utilizadas pelo método proposto, bem como a justificativa para sua escolha. Foram realizados experimentos comparando três técnicas de sumarização que estão implementadas na linguagem de programação Python. Essas técnicas são: LSA (*Latent Semantic Analysis*), SumBasic e Edmundson. Os experimentos foram executados através da seleção aleatória de dez documentos texto que compõe a base de dados utilizado em um dos estudos de casos abordados nesta pesquisa. Para cada documento foi construído um sumário a partir da avaliação de uma pessoa (regra ouro) e em seguida o mesmo documento é submetido as técnicas de sumarização automática. O resultado de cada sumarização é comparado a respectiva regra ouro. Desta maneira, o algoritmo LSA apresentou os melhores sumários em relação a regra ouro estabelecida.

Um experimento similar foi realizado para selecionar a ferramenta de anotação

de entidades (NER/NED). O experimento contou com cinco ferramentas de anotação: Babelfy[8], Dbpedia Spotlight[9], Google Cloud Natural Language (Google CNL), Open Calais e IBM Watson Natural Language Understanding (IBM Watson NLU). Alguns documentos textos tiveram suas entidades anotadas manualmente e automaticamente pelas ferramentas. Em seguida, foi utilizado dois critérios para comparar as ferramentas, cobertura (i.e. o número de entidades anotadas pela ferramenta) e precisão (i.e. quantas entidades foram anotadas corretamente). Nesses experimentos destacaram-se o Google CNL e o IBM Watson NLU, sendo a ferramenta do Google selecionada ao final. A ferramenta de anotação selecionada conta com um módulo de anáise de sentimento e uma base de conhecimento própria que também são utilizados nessa pesquisa.

A obtenção das redes semânticas ocorre através da técnica de seleção a partir de uma grande base de conhecimento. A base de conhecimento utilizada foi o Google KB e as seleções ocorreram através do tipo de entidade e o papel que ela exerce, além da relação entre as entidades (e.g. entidades do tipo pessoa com papel de político, entidades do tipo organização com papel de partido político, e relação de filiados), em determinado período de tempo.

A influência (peso) de cada entidade pertencente a rede semântica foi obtida através do uso de um algoritmo genético. A correlação utilizada foi a correlação de Pearson. Por fim, quando existe correlação entre as variáveis é aplicado a regressão linear simples, caso contrário é utilizado o classificador OneR. O uso desse classificador acontece devido a sua aplicação em trabalhos relacionados.

[8] http://babelfy.org/
[9] https://www.dbpedia-spotlight.org/

3 TRABALHOS RELACIONADOS

Este capítulo discute trabalhos relacionados ao tema dessa pesquisa que foram coletados em um processo de revisão sistemática da literatura. As bases consultadas foram: IEEE Xplore, ICM Digital Library, Scopus, Scielo, Springer, BDBComp e Google Scholar. As expressões de busca usadas incluíram as palavras-chaves: *menções (mention, reference, quote, citation)*, *entidades (entity, related entities)*, *sumarização (summarization)*, *análise de sentimentos (sentiment analysis, opinion mining, emotion AI)*, *desempenho (entity performance, agent performance, performance, rating, evaluation, value)* e *correlação (correlation, correlativity, influence)*. O período de publicação considerado foi superior ou igual ao ano de 2012. Na primeira rodada de buscas foram identificados 973 artigos. Em seguida foram verificados os títulos e resumos desses artigos, para selecionar dentre eles os mais relacionados ao tema da pesquisa. Ao final desse processo foram selecionados 35 artigos relacionados, dentre os quais se destacam os citados neste capítulo.

Este capítulo está dividido em três seções, que podem ser descritas da seguinte maneira. A Seção 3.1 discute os principais trabalhos que calculam métricas como incidências de menções a entidades em texto, algumas dessas métricas levando em conta os sentimentos associados, e que investigam correlações entre tais métricas e alguma medida real de desempenho das entidades. A Seção 3.2 analisa pesquisas que predizem a medida real de desempenho usando as métricas produzidas a partir das características extraídas de textos. Por fim, a Seção 3.3 compara os trabalhos discutidos nas seções anteriores com o método *5-Ions*, proposto em nosso trabalho.

3.1 MÉTRICAS DERIVADAS DE MENÇÕES A ENTIDADES EM TEXTO E SUA CORRELAÇÃO COM DESEMPENHO REAL DE ENTIDADES

Vários trabalhos da literatura exploram menções a entidades encontradas em textos para análise de informação neles contida. Alguns desses trabalhos usam técnicas de *Business Intelligence* (BI) para analisar as distribuições das menções a entidades em textos usando hierarquias de tempo, espaço e (classes das) respectivas entidades (NEBOT; BERLANGA, 2012; FRANCIA *et al.*, 2014; KUMAR *et al.*, 2013; FILETO *et al.*, 2014; SACENTI *et al.*, 2015; VILLANUEVA *et al.*, 2016; PEREIRA *et al.*, 2018).

Tais hierarquias podem ser obtidas de redes semânticas existentes, tais como grafos de conhecimento, por exemplo. Esses trabalhos consideram processos de extração, transformação e carga de dados que incluem tarefas de anotação semântica de textos. Os dados semanticamente enriquecidos alimentam bases de dados dimensionais para suportar a análise da variação de métricas como popularidade (incidência) de menções a certas (classes de) entidades ao longo do tempo, em regiões e locais

de interesse. Esses trabalhos também estão entre os poucos que aplicam ferramentas de PLN para extrair e analisar informação de texto, ao invés de utilizar técnicas simples (e.g. baseadas em dicionários de nomes de superfície) para reconhecer as menções nos textos. Entretanto, poucos trabalhos têm tentado avaliar correlações entre métricas derivadas de informações extraídas de textos e medidas de desempenho real, e/ou predizer o desempenho futuro de entidades a partir de suas menções atuais e históricas em textos.

A pesquisa (AHMED; SKORIC, 2014) analisa o impacto de *tweets* em Inglês nas eleições de 2013 do Paquistão. (AHMED; SKORIC) utilizam dados estatísticos oficiais da eleição como a medida real de desempenho, além dos *tweets* de onde são extraídas as informações para análise. *Tweets* com texto em Urdu, outro idioma oficial do Paquistão, são ignoradas devido à falta de recursos computacionais para analisar textos nesse idioma. Cada *tweet* restante (usualmente em inglês) é classificado quanto ao seu autor e quanto ao conteúdo do seu texto. Os autores são classificados em: *pessoa*, *político*, *jornal* ou *outros*. Já os textos dos *tweets* são classificados em: *atualização de campanha*, *promoção*, *crítica*, *chamada de voto*, *notícias políticas*, *detalhes de partido*, *outras notícias* ou *outros*. Após as classificações, os *tweets* são ordenados de acordo com a data de sua publicação e a frequência das suas classes são comparadas com as estatísticas oficiais no mesmo período. Os autores afirmam que os resultados são inconclusivos, porém promissores para provar o impacto da mídia social nas eleições do Paquistão.

A pesquisa (FINK *et al.*, 2013), por sua vez, analisa o impacto dos sentimentos associados aos *tweets* nas eleições da Nigéria no ano de 2011 e na aprovação do atual presidente. Os candidatos analisados são: Jonathan (candidato a reeleição), Buhari, Ribadu e Shekarau. Os autores selecionam apenas *tweets* publicados na Nigéria entre janeiro e abril de 2011 e que mencionam algum desses candidatos. As eventuais ambiguidades encontradas ao anotar as menções às entidades são resolvidas com a aplicação de alguns testes estatísticos. Os textos são pré-processados para resolver problemas como abreviações e uso de linguagem coloquial da internet. Em seguida, alguns classificadores de sentimento são utilizados para identificar a polaridade dos textos. Tanto a frequência das menções quanto a frequência dos sentimentos anotados para as entidades são correlacionados (correlação de Spearman e correlação de Pearson, respectivamente) com a quantidade de votos recebidos. Segundo essas medidas apenas a entidade Jonathan apresenta valor significativo de correlação (coeficiente de Pearson de .86) entre a frequência de sentimento positivo e votos recebidos. As demais análises não apresentam resultados significativos.

3.2 PREDIÇÃO DE DESEMPENHO DE ENTIDADES

O método proposto em (RAMTEKE *et al.*, 2016) prediz o vencedor de uma eleição classificando o sentimento das menções aos candidatos e aos seus partidos políticos em *tweets*. Como estudo de caso, (RAMTEKE *et al.*) analisam as eleições dos Estados Unidos de 2016. O trabalho compara a métrica F-1 Score de um classificador naïve Bayes com um classificador SVM (*Support Vector Machine*). Esses classificadores são treinados com uma base de dados criada manualmente para predizer o sentimento contido nas publicações. Por fim, o método considera vencedor da eleição o candidato com a maior razão entre a frequência de sentimentos positivos e a frequência de sentimentos negativos.

Por sua vez, o artigo (TUMITAN; BECKER, 2014) utiliza séries temporais de sentimentos extraídos automaticamente de comentários online de usuários em artigos de notícias. As séries temporais de sentimentos permitem predizer a variação da intenção de voto de duas eleições no Brasil. O processo de anotação de entidades utiliza um dicionário de palavras com possíveis nomes, sobrenomes e apelidos dos principais candidatos as eleições. As séries temporais de sentimentos são utilizadas para calcular métricas de desempenho para cada candidato e seu respectivo partido político. O resultado do cálculo da métrica é utilizado para predizer a variação da intenção de voto através da aplicação do algoritmo de classificação OneR. A acurácia da predição é usada para validar o método proposto. A máxima acurácia obtida pelos autores foi de 70% para classificações binárias e 56% para classificadores ternários.

3.3 ANÁLISE COMPARATIVA

O Quadro 2 sintetiza os trabalhos apresentados neste capítulo. A primeira coluna refere-se aos trabalhos analisados, já as três colunas seguintes apresentam as características extraídas em texto que são abordadas nesta pesquisa (*Sumarização*, *NER e NED* e *AS* (Análise de Sentimento), respectivamente). A coluna *RS* (Rede Semântica) indica se os trabalhos consideram a influência de entidades relacionadas em suas análises. Por fim, a coluna *Investigação* informa se o trabalho busca correlacionar e/ou predizer alguma medida de desempenho. A última linha foi adicionada para representar a proposta deste trabalho, o método *5-Ions*.

O tamanho dos documentos analisados torna desnecessário a aplicação de sumarização de texto nos trabalhos analisados (e.g., *tweets* limitados a 144 caracteres, comentário de notícias online com poucas palavras). Todos os trabalhos que realizam alguma investigação de desempenho utilizam dicionários de dados para mapear os nomes de superfície e aplicam métodos bastantes rudimentares para reconhecer entidade nomeada em texto. Para tal, utilizamos ferramentas de anotação que realizam o reconhecimento e a desambiguação das entidades nomeadas. Já a análise de sen-

Quadro 2 – Quadro comparativo dos trabalhos relacionados.

Trabalho	Sumarização	NER e NED	AS	RS	Investigação
(NEBOT; BERLANGA, 2012)		Ferramenta de anotação		X	
(FRANCIA *et al.*, 2014)		Ferramenta de anotação		X	
(KUMAR *et al.*, 2013)		Ferramenta de anotação		X	
(FILETO *et al.*, 2014)		Ferramenta de anotação		X	
(SACENTI *et al.*, 2015)		Ferramenta de anotação		X	
(VILLANUEVA *et al.*, 2016)		Ferramenta de anotação		X	
(PEREIRA *et al.*, 2018)		Ferramenta de anotação		X	
(AHMED; SKORIC, 2014)		dicionário			Correlação
(FINK *et al.*, 2013)		dicionário	X		Correlação
(RAMTEKE *et al.*, 2016)		dicionário	X		Predição
(TUMITAN; BECKER, 2014)		dicionário	X		Predição
Nossa proposta (5-Ions)	**X**	**Ferramenta de anotação**	**X**	**X**	**Correlação e Predição**

Fonte: elaborado pelo autor.

timentos tornou-se algo natural nas investigações de desempenho das entidades. De fato, não foi encontrado nenhum trabalho na literatura, até hoje, que explore relações semânticas entre entidades em uma rede de relacionamento para estimar o desempenho de uma entidade alvo, com base em entidades semanticamente relacionadas. Portanto, pelo que sabemos, este trabalho é o primeiro a utilizar (Capítulo 4) e comparar os resultados obtidos analisando apenas as menções a uma entidade alvo versus análises que mencionam entidades semanticamente relacionadas (Capítulos 5 e 6).

4 PROPOSTA: *5-IONS*

Este capítulo apresenta, em detalhes, o método proposto para estimar medidas reais de desempenho de entidades utilizando ferramentas, atuais e disponíveis, de processamento de linguagem natural (PLN) para extrair características de texto, ou seja, entidades nomeadas e seu sentimento associado, e explorar as relações semânticas dessas entidades. O método utiliza as características extraídas de textos e as relações semânticas das entidades para compor métricas de desempenho que permitam estimar o desempenho real de uma entidade alvo. Esse método é chamado de *5-Ions* porque contém 5 etapas nomeadas com terminações em "ion", são elas: *Pre-processing & Summarization*, *Annotation*, *Consolidated Calculation*, *Correlation* e *Prediction*. Cada etapa envolve uma ou mais tarefas de processamento de dados, algumas delas suportadas por ferramentas de PLN.

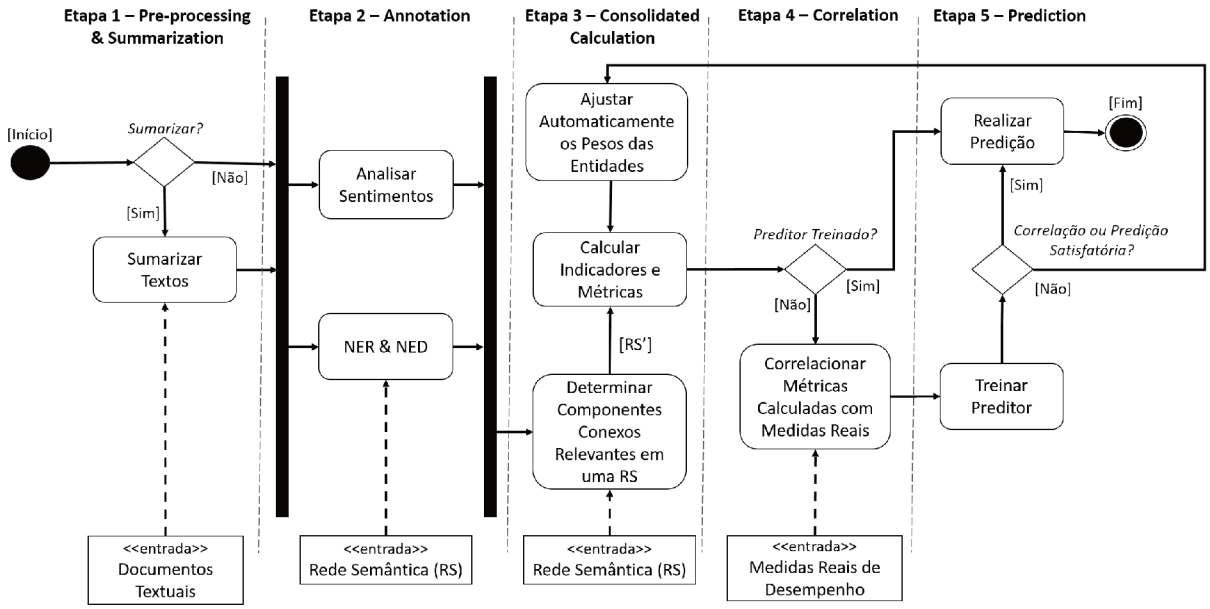
A seguir, descrevemos em detalhes alguns aspectos do método proposto. A Seção 4.1 descreve o processo geral do método *5-Ions*, o qual permite adaptações para necessidades específicas, alterando fonte de dados, ferramentas para realização das tarefas, e para aplicar ou não alguma tarefa. A Seção 4.2 descreve os indicadores e as métricas de desempenho consolidadas para entidades semanticamente relacionadas. Finalmente, a Seção 4.3 detalha o algoritmo genético utilizado para ajustar os pesos das entidades envolvidas no cálculo das métricas de desempenho consolidadas (para entidades relacionadas semanticamente) de tal forma que maximiza sua correlação, ou a acurácia das predições, com as medidas reais de desempenho.

4.1 PROCESSO GERAL

A Figura 4 apresenta o processo geral do método *5-Ions*. O método possui três entradas ao longo do processo, são elas: *documentos textuais*, *rede semântica (RS)* e *medidas reais de desempenho*. Os documentos textuais contêm menções, e sentimento associado as menções, que são identificadas e vinculadas as entidades descritas na RS. Esses documentos possuem data e hora do momento da publicação, o que permite comparar as métricas de desempenho derivadas desses documentos com as medidas reais de desempenho oriundas de outras fontes no mesmo período (e.g. intenção de voto para um candidato ou partido político). Mesmo quando as métricas de desempenho não apresentam correlações significativas com a medida real de desempenho elas ainda podem ser utilizadas para predizer, ao menos, a variação das medidas reais nos tempos subsequentes ao das suas publicações. Para tal, um subconjunto das medidas reais de desempenho é utilizado para treinar os modelos de aprendizado de máquina para predição, enquanto outro subconjunto é usado para avaliar as predições geradas.

O processo *5-Ions* tem início na etapa intitulada *Pre-processing & Summari-*

Figura 4 – O Processo Geral do Método *5-Ions*.



Fonte: elaborado pelo autor.

zation (Pré-processamento & Sumarização). A primeira ação do processo é decidir se é necessário sumarizar os documentos textuais. Eventualmente, uma tarefa para pré-processar os textos pode ser necessária para resolver ruídos e outros problemas presentes em alguns textos. Por exemplo, publicações em mídia social, usualmente, possuem erros de digitação, gírias, acrônimos, abreviações, entre outros problemas, que necessitam a aplicação de técnicas de normalização de palavras (BONTCHEVA; ROUT, 2012). A tarefa *Sumarizar Textos* diminui a quantidade dos dados que são processados nas etapas seguintes.

A etapa 2 compreende a execução de duas tarefas de anotação (*Annotation*), *NER & NED* e *Analisar Sentimentos*. A execução dessas tarefas, na prática, podem estar entrelaçadas. Ambas as atividades dependem de técnicas de PLN, como análise léxica (*tokenization*) e *Pos-Tagging*. De fato, como mencionado no Capítulo 2.2 algumas ferramentas executam ambas as tarefas de uma vez. Seja $D = \{d_1, \ldots, d_n\}$ um conjunto de $n > 0$ documentos de texto, sumarizados ou não. Cada documento $d_i \in D$ ($1 \leq i \leq n$) é uma tripla $d_i = \langle idd, timestamp, text \rangle$, onde *idd* corresponde ao identificador do documento, *timestamp* corresponde a data e hora da publicação do documento e *text* corresponde o conteúdo textual. São utilizadas algumas ferramentas de PLN no estado da arte, como o *Google Cloud Natural Language*, que reconhecem automaticamente as menções contidas nos textos do conjunto D, calcula o sentimento associado a essas menções e vincula cada menção ao seu respectivo recurso na rede semântica $RS(V, E)$. As menções anotadas podem incluir correferências ou não, de acordo com as necessidades da aplicação. A saída da etapa 2 é um conjunto A de anotações

semânticas das entidades encontradas nos documentos em D. Cada anotação $a \in A$ é uma tupla $a = \langle d, m, v, s \rangle$, onde $d \in D$ é um documento, m é uma menção encontrada em d, $v \in V$ é a entidade da RS a qual m está vinculada, e $s \in [-1, 1]$ é o sentimento associado a menção m.

A primeira tarefa executada durante a etapa 3 é *Determinar Componentes Conexos Relevantes em uma RS*. Essa tarefa produz um extrato $RS'(V', E') = \{C_1, \ldots, C_m\}$ de uma rede semântica $RS(V, E)$, onde $V' \subseteq V$, $E' \subseteq E$, abstraindo os sentidos das arestas em E'. O grafo $RS'(V', E')$ é dividido em um conjunto de componentes conexos $\{C_1, \ldots, C_m\}$, como ilustrado no exemplo da Figura 1 (Capítulo 2.1). Essa atividade é necessária quando a RS utilizada para anotar as menções às entidades nos textos é muito grande ou não explicita os componentes conexos. Alguns métodos para determinar componentes conexos relevantes em uma RS são apresentados no Capítulo 2.

A próxima tarefa da etapa 3 a ser executada é *Calcular Indicadores e Métricas*. Dado o conjunto A das anotações dos documentos em D, gerado na etapa 2, e o extrato $RS'(V', E') = \{C_1, \ldots, C_m\}$ da rede semântica $RS(V, E)$ (utilizada para gera as anotações em A), contendo as entidades consideradas relevantes para as análises organizadas em componentes conexos $C_1, \ldots, C_m$ é possível então calcular os indicadores e métricas de desempenho consolidadas para entidades semanticamente relacionadas (pertencentes ao mesmo componente conexo $C_j \in RS'$ $(1 \leq j \leq m)$) nos períodos de tempo das publicação dos documentos (em conformidade com as medidas de desempenho real). As fórmulas utilizadas para calcular os indicadores e métricas consolidadas são detalhadas na seção seguinte.

A última atividade da terceira etapa, *Ajustar Automaticamente os Pesos das Entidades*, define automaticamente pesos para as entidades pertencentes a RS' que compõe o cálculo dos indicadores e métricas consolidadas. A Seção 4.3 detalha como essa tarefa é executada automaticamente através do uso de um algoritmo genético.

Durante a única tarefa da etapa 4, *Correlacionar Métricas Calculadas com Medidas Reais*, os indicadores e as métricas de desempenho, produzidos na etapa anterior, são correlacionados, ao longo do tempo, com as medidas reais de desempenho, coletadas em outras fontes. Os indicadores, as métricas e as medidas reais são considerados séries temporais. A comparação entre as séries pode coincidir perfeitamente no tempo ou pode exigir um deslocamento temporal de alguma das séries (e.g. uma métrica pode ser correlacionada com uma medida real de um tempo subsequente ou vice-versa). Coeficientes de correlação distintos podem ser empregados. Em alguns casos, a existência de correlação significativa entre as séries temporais pode habilitar o uso de modelos estatísticos para realizar a predição do desempenho. Quando não há interesse em avaliar a correlação (e.g. já existe um preditor treinado ou as correlações não são necessárias para o treinamento) o processo pode pular a etapa 4 e seguir para a etapa 5.

Finalmente, durante a etapa 5, a estratégia de escolha para a tarefa *Treinar Preditor* pode depender dos resultados das correlações. Se uma correlação significativa for encontrada entre as métricas e as medidas reais de desempenho, algum modelo estatístico poderá ser aplicado para realizar a predição da medida real. De outra forma, algoritmos de aprendizado de máquina são utilizados para treinar o preditor. A tarefa *Realizar Predição* trabalha de acordo. Se o preditor é baseado em correlações estatísticas, as métricas de desempenho são combinadas para prever as medidas de desempenho. Caso contrário, um preditor baseado em aprendizado de máquina é utilizado para predizer certos tipos de variações (e.g. aumenta, diminui ou permanece estável) da medida real de desempenho.

4.2 INDICADORES E MÉTRICAS CONSOLIDADAS PARA ENTIDADES RELACIONADAS

Os indicadores e métricas de desempenho consolidadas para entidades semanticamente relacionadas requerem um conjunto de documentos D com as menções as entidades identificadas e vinculadas as entidades de uma rede semântica $RS(V,E)$, e os sentimentos associados a essas menções determinadas. As menções às entidades anotadas podem ser representadas por um conjunto de anotações semânticas A, conjunto esse resultante da etapa 2 do processo geral $5-Ions$. Os indicadores e métricas consolidadas necessitam de um conjunto relevante de componentes conexos $RS'(V',E') = \{C_1,\ldots,C_m\}$, fornecidos pela tarefa *Determinar Componentes Conexos Relevantes em uma RS*, de tal modo que $V' \subseteq V$, e a cada aresta não direcionada $e'(v_i,v_j) \in E'$ há uma aresta direcionada $e(v_i,v_j) \in E$. Além disso, pelo menos os conjuntos dos vértices dos componentes conexos $C_1,\ldots,C_m$ devem constituir uma partição do conjunto de vértices V' da $RS'(V',E')$. Mais formalmente, para $1 \leq i,j \leq m$ e $i \neq j$:

$$\forall\, C_i(V_i,E_i), C_j(V_j,E_j) \in RS'(V',E') : V_i \cap V_j = \emptyset \tag{1}$$

$$V' = \bigcup_{i=1}^{m} V_i \tag{2}$$

Então, é possível calcular os seguintes indicadores de desempenho consolidados baseados nas menções às entidades e suas respectivas polaridades do conjunto A das menções de um conjunto de documentos D, para um conjunto de entidades em um componente conexo $C_j(V_j,E_j)$ da $RS'(V',E')$, e um dado intervalo de tempo $\rho[t_{start},t_{end}]$, do seguinte modo:

$$pos_{V_j,\rho} = \sum_{v \in V_j,\ \ a \in A \,\wedge\, a.v=v \,\wedge\, a.d.t \in \rho \,\wedge\, a.s>0} a.s \tag{3}$$

$$neg_{V_j,\rho} = \sum_{v \in V_j,\ \ a \in A \ \wedge\ a.v=v \ \wedge\ a.d.t \in \rho \ \wedge\ a.s<0} |a.s| \tag{4}$$

Em outras palavras, o indicador de desempenho consolidado $pos_{V_j,\rho}$ ($neg V_j, \rho$) é a soma dos sentimentos positivos (negativos) $a.s$ (módulo $|a.s|$) de todas as menções anotadas $a \in A$ dos documentos em D de tal forma que as menções anotadas $a.m$ estão ligadas a uma entidade $a.v \in V_j$ e seu sentimento associado é $a.s \in (0,1]$ ($a.s \in [-1,0)$). Esses indicadores de desempenho também são consolidados para as anotações cuja a data e hora de publicação dos documentos $a.d.t$ estão contidos no intervalo de tempo $\rho[t_{start}, t_{end}]$, sendo t_{start} seu tempo de início e t_{end} seu tempo de término. Note que essas fórmulas generalizam os indicadores de desempenho empregados em (TUMITAN; BECKER, 2014) para uma determinada entidade em particular v e um instante determinado t que pode ser apresentado da seguinte forma:

$$pos_{v,t} = \sum_{a \in A \ \wedge\ a.v=v \ \wedge\ a.d.t=t \ \wedge\ a.s>0} a.s \tag{5}$$

$$neg_{v,t} = \sum_{a \in A \ \wedge\ a.v=v \ \wedge\ a.d.t=t \ \wedge\ a.s<0} |a.s| \tag{6}$$

Analogamente, as métricas de desempenho empregadas em (TUMITAN; BECKER, 2014) podem ser consolidadas, como especificado no Quadro 3, para um conjunto de vértices V_j de um componente conexo $C_j(V_j, E_j)$ da $RS'(V', E')$. Note que as métricas f_5, f_6 e f_7 consolidam os indicadores de desempenho em seus denominadores para todos V' da $RS'(V', E')$. Os pesos $w_v \in \mathbb{R}$ são utilizados para dar uma importante distinção para cada entidade v do componente conexo V_j. Esses são apenas alguns exemplos de métricas consolidadas que podem ser empregadas para investigar as correlações com as medidas reais de desempenho, e para predizer tais medidas ou suas variações.

4.3 AJUSTE DE PESO DAS ENTIDADES

Cada entidade dos componentes conexos da rede semântica (RS') analisada possui uma influência (peso) em relação a alguma entidade alvo. Porém a definição desse peso não é uma tarefa trivial. Esses pesos podem ser definidos através de modelos estatísticos (e.g. regressão linear múltipla) ou através do uso de algoritmos de busca por soluções e de otimização (e.g. algoritmo genético). A utilização de modelos estatísticos muitas vezes requer o atendimento de alguns critérios, por exemplo, o fato da variável (medida real de desempenho) ser continua, discreta ou categórica. Já para algoritmos de busca por soluções o tipo da variável não é um impeditivo para sua utilização. Por esse motivo, esse trabalho utiliza algoritmos de busca, mais especificamente o algoritmo genético.

Quadro 3 – Métricas de desempenho consolidadas para entidades em um componente conexo $C_j(V_j, E_j)$ da *RS* e período de tempo ρ.

Métrica	Fórmula	Descrição
f_1	$f_1(V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} pos_{v,t}}{\sum_{t\epsilon\rho} neg_{v,t}}$	Razão do sentimento positivo em relação a um conjunto de entidades relacionadas sobre o sentimento negativo das mesmas entidades.
f_2	$f_2(V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} pos_{v,t}}{\sum_{t\epsilon\rho} \langle pos_{v,t}+neg_{v,t}\rangle}$	Razão do sentimento positivo em relação a um conjunto de entidades relacionadas sobre o sentimento total das mesmas entidades.
f_3	$f_3(V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} neg_{v,t}}{\sum_{t\epsilon\rho} \langle pos_{v,t}+neg_{v,t}\rangle}$	Razão do sentimento negativo em relação a um conjunto de entidades relacionadas sobre o sentimento total das mesmas entidades.
f_4	$f_4(V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} \langle pos_{v,t}-neg_{v,t}\rangle}{\sum_{t\epsilon\rho} \langle pos_{v,t}+neg_{v,t}\rangle}$	Razão da diferença entre o sentimento positivo e negativo em relação a um conjunto de entidades relacionadas sobre o sentimento total das mesmas entidades.
f_5	$f_5(V', V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} pos_{v,t}}{\sum_{q\epsilon V'} \sum_{t\epsilon\rho} pos_{q,t}}$	Razão do sentimento positivo em relação a um conjunto de entidades relacionadas sobre o sentimento positivo total (todas as entidades, independente de relação).
f_6	$f_6(V', V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} neg_{v,t}}{\sum_{q\epsilon V'} \sum_{t\epsilon\rho} neg_{q,t}}$	Razão do sentimento negativo em relação a um conjunto de entidades relacionadas sobre o sentimento negativo total (todas as entidades, independente de relação).
f_7	$f_7(V', V_j,\rho) = \sum_{v \in V_j} w_v \cdot \frac{\sum_{t\epsilon\rho} men_{v,t}}{\sum_{q\epsilon V'} \sum_{t\epsilon\rho} men_{q,t}}$	Razão das menções em relação a um conjunto de entidades relacionadas sobre o total de menções (todas as entidades, independente de relação).

Fonte: Adaptado de (TUMITAN; BECKER, 2014).

Os pesos w_v das métricas de desempenho consolidadas (Quadro 3) são ajustados automaticamente, por um Algoritmo Genético (AG), para maximizar as associações (i.e., coeficientes de correlação, acurácias das predições) entre as métricas consolidadas e as medidas reais de desempenho (RUSSELL; NORVIG, 2003). Cada peso $w_v \in \mathbb{R}$ é considerado um elemento representado por um vetor de dígitos no conjunto $\{0,1,\ldots,9\}$ com uma posição adicional para representar o conjunto dos sinais $\{+,-\}$. Assim, o peso –9.45678 é representado por um vetor de tamanho 7. Os indivíduos da população no AG são vetores de elementos. Cada vetor tem o número de elementos

do conjunto V_j.

A função *fitness* do AG depende da associação que se deseja maximizar (i.e. correlação ou predição). É necessário informar o valor buscado (e.g., correlação superior a .75, acurácia superior a .8) e o número máximo de gerações para garantir a parada do AG. A função *fitness* é aplicada aos indivíduos da população para alguma métrica consolidada (Quadro 3), previamente definida. A métrica resultante é associada a medida real de desempenho e o coeficiente da associação resultante representa o *fitness* do indivíduo.

O cruzamento de cromossomos dos indivíduos da população ocorre aleatoriamente. A cada novo cruzamento é definido um novo ponto de corte. Esse ponto de corte é maior que zero e menor que o tamanho do cromossomo. Cada par de indivíduo (pais) selecionado têm seus elementos combinados resultando em um novo indivíduo (filho). Cada elemento do filho tem genes de um dos pais até o ponto de corte e os demais genes são oriundos do outro pai. A cada novo cruzamento pode ocorrer uma mutação nos elementos do filho. Se uma mutação ocorrer um gene cromossômico é substituído por um novo gene aleatório, i.e. algum dígito do vetor representante do elemento é substituído por um novo dígito.

5 EXPERIMENTOS

Este capítulo descreve os experimentos realizados para testar o método proposto. O principal objetivo desses experimentos é comparar a qualidade das correlações e das predições obtidas através das métricas de desempenho calculadas para cada entidade a partir de suas menções e sentimentos em textos com aquelas obtidas com métricas análogas, consolidadas para entidades semanticamente relacionadas em uma rede semântica (RS), que são propostas neste trabalho (Quadro 3). O uso de um algoritmo genético para compor métricas consolidadas, ajustando a influência (peso) das entidades que compõe o componente conexo que contém a entidade alvo, permite obter melhores correlações com a medida real desempenho de entidades e melhores predições deste. Essas melhorias são avaliadas em dois estudos de caso, em política Alemã e oscilações da cotação do real, respectivamente. Os experimentos também avaliam o impacto de uma técnica selecionada de sumarização de texto no tempo de execução e nos resultados obtidos pelo método *5-Ions*, proposto neste trabalho, além do impacto de correferências detectadas por uma ferramenta de anotação semântica de textos.

A Seção 5.1 descreve os dois estudos de caso que são abordados nos experimentos. O primeiro estudo de caso apresenta escassez temporal e baixa confiabilidade das medidas reais, i.e., as medidas reais de desempenho são pesquisas de opinião com poucas publicações ao longo do tempo. Já o segundo estudo de caso ostenta grande volume de publicações das medidas reais, porém os textos analisados possuem poucas menções a entidade alvo do estudo. A Seção 5.2 detalha as configurações do método e dos componentes computacionais empregadas nos experimentos por estudo de caso.

5.1 ESTUDOS DE CASO

Dois estudos de caso são considerados neste trabalho. O primeiro (Seção 5.1.1) aborda as eleições Alemãs de 2017, vencida pela união partidária CDU/CSU da chanceler Angela Merkel. O segundo estudo de caso (Seção 5.1.2) analisa a flutuação do Real, a moeda oficial do Brasil, durante a instabilidade política/econômica do país dentre os anos de 2014 e 2015.

5.1.1 Intenção de Voto

Esse estudo de caso avalia a intenção de voto nos principais candidatos e seus respectivos partidos políticos ao longo da campanha eleitoral Alemã de 2017. A escolha pela política Alemã acontece devido à alta disponibilidade de artigos de notícias e pesquisas de opinião referentes a campanha na *Web*. Esse estudo permite analisar o

impacto das menções as entidades relacionadas no desempenho de algumas entidades alvos, condição que não foi verificada em outros trabalhos. Por exemplo, como os candidatos mais populares influenciam o desempenho dos seus partidos/alianças, e vice-versa.

Somente os partidos mais votados são considerados nesse experimento, são eles: Partido Social-Democrata da Alemanha (*Sozialdemokratische Partei Deutschlands* - SPD); aliança entre a União Democrata-Cristã e sua contrapartida Bávara (*Christlich-Demokratische Union Deutschlands* e *Christlich-Soziale Union in Bayern* - CDUCSU); Esquerda (*Die Linke* - LINKE); Aliança 90 e os Verdes (*Bündnis 90Die Grünen* - GRUNE); Partido Democrático Liberal (*Freie Demokratische Partei* - FDP); e Alternativa para a Alemanha (*Alternative für Deutschland* - AfD).

5.1.2 Flutuação da Moeda

Esse estudo de caso investiga a flutuação da moeda Brasileira, o Real, com base nas menções às entidades que usualmente a influenciam. Dentre essas entidades temos o legislativo federal (e.g. senadores), membros do governo federal (e.g., presidente, ministros), bem como várias instituições, públicas ou privadas (e.g., bancos, companhias importantes para setores da economia Brasileira como energia e mineração).

O processo de obtenção da rede semântica ocorre através de seleções sucessivas de entidades do tipo organização e entidades do tipo pessoa com diversos papeis, em determinado período de tempo, e das relações entre as entidades do tipo pessoa e do tipo organização. Alguns desses papeis são descritas em (ARVATE, 2004; ROSSI, 2015). Dentre elas temos membros e instituições do poder executivo que incluem o presidente da república, alguns ex-presidentes e alguns ministros (e respectivos ministérios), totalizando 10 entidades. Além disso, menções aos 81 senadores e outras 18 instituições e representantes, principalmente empresas, são consideradas.

5.2 CONFIGURAÇÕES DOS EXPERIMENTOS

As configurações dos experimentos estão divididas em seções. A Seção 5.2.1 destaca os locais de busca dos textos e das medidas reais de desempenho de cada estudo de caso, bem como o critério utilizado para obter esses dados. A Seção 5.2.2 apresenta as ferramentas, técnicas e algoritmos utilizados nos experimentos, além das suas configurações. Por fim, a Seção 5.2.3 informa os recursos de hardware e software ao qual os experimentos são submetidos.

5.2.1 Bases de Dados

Os documentos textos para o estudo de caso da intenção de voto nas eleições Alemã de 2017 foram obtidos na Deutsche Welle (DW)[1]. Cada texto capturado contém, pelo menos, o nome ou acrônimos de algum dos principais partidos ou políticos alemães. A aliança CDU/CSU é considerada uma única entidade (partido político). Por se tratar de uma mídia especializada, as publicações do DW tendem a apresentar baixo ruído. A escolha desse tipo de mídia é uma tentativa de maximizar os resultados das anotações e facilitar a descoberta da RS. Entretanto, o método proposto pode ser usado com qualquer tipo de documento textual (e.g., comentário de notícias, *tweets*).

São capturados 5117 documentos publicados pela DW, escritos em Alemão, entre as datas de 25 de dezembro de 2016 e 25 de setembro de 2017. Os textos são traduzidos para o idioma Inglês através da ferramenta Google Translate[2]. A tradução automática é necessária para que ocorra a medição do sentimento no contexto de entidade. Essa funcionalidade, disponível em algumas ferramentas (e.g., Watson NLU, Google CNL), exige que os textos estejam escritos em Inglês. É sabido que isso pode impactar nos resultados, entretanto, trabalhos anteriores mostram que esse impacto não é determinante (CIRQUEIRA *et al.*, 2017). Além disso, experimentos preliminares com tradução de texto mostraram bons resultados com baixa perda semântica.

A medida de desempenho real, para esse estudo de caso, é obtida através dos institutos *TNS Emnid GmbH & Co. KG (Emnid)* e *Forschungsgruppe Wahlen*. O Emnid divulgou 38 pesquisas de intenção de voto em 2017, até a data das eleições. A *Forschungsgruppe Wahlen*, por sua vez, divulgou 17 pesquisas durante o mesmo período. A última pesquisa do instituto *Forschungsgruppe Wahlen* previu corretamente a ordem dos cinco partidos mais votados. Já a última pesquisa do Emnid acertou a ordem dos três partidos mais votados, porém, previu que o partido FDP seria o quarto partido mais votado seguido pelo partido LINKE, em quinto lugar. Entretanto, esses partidos obtiveram ordem inversa no resultado oficial, i.e., LINKE foi o quarto colocado e o FDP o quinto partido mais votado.

Já para o estudo de caso de flutuação da moeda, os documentos textos são coletados no Valor Econômico[3], um jornal Brasileiro com foco em economia, finanças e negócios. Ao todo são capturados 20435 artigos, publicados entre as datas de 01 de abril de 2015 e 31 de dezembro de 2016 e que mencionam alguma das entidades apresentadas na seção anterior. Todos os documentos são originalmente escritos em Português.

A medida de desempenho real considerada, nesse estudo de caso, é a taxa de câmbio do Real Brasileiro em relação ao Dólar Americano. A taxa de câmbio é

[1] http://www.dw.com/
[2] https://translate.google.com/
[3] https://www.valor.com.br/

disponibilizada pelo Banco Central Brasileiro (BACEN)[4] durante o mesmo período das publicações dos documentos pelo Valor Econômico. Durante esse período o BACEN publicou 440 cotações da moeda, no qual 1 dólar varia de 2,89 a 4,19 reais.

5.2.2 Ferramentas, Algoritmos e Técnicas

A atividade *Sumarizar Textos* (Capítulo 4.1) utiliza o algoritmo de sumarização LSA com fator de compreensão de 30%, i.e., o documento original é reduzido a 30% do seu tamanho original. Por ser opcional, essa atividade só é executada para o estudo de caso de intenção de voto.

As atividades da *Etapa 2 (Annotation)* são executadas com auxílio da ferramenta Google Cloud Natural Language (Google CNL) em sua versão de testes. A RS utilizada por essa ferramenta é a Google Knowledge Graph (Google KG). No estudo de caso de intenção de voto, a atividade *Determinar Componentes Conexos Relevantes em uma RS* extrai um subgrafo da Google KG com os principais partidos Alemão e os políticos filiados mais populares desses partidos. A RS extraída contém apenas 6 partidos (considerando a aliança CDU/CSU como um único partido), e 37 políticos, cada um vinculado a seu respectivo partido. Já para o estudo de caso da flutuação da moeda, a RS extraída contém 10 organizações (dentre ministérios, empresas privadas e empresas públicas) e 38 pessoas, incluindo 3 ex-presidentes da república.

Visando demonstrar a versatilidade do método *5-Ions*, mantém-=se o idioma original do estudo de caso da flutuação da moeda. A Google CNL determina o sentimento associado as menções, porém apenas para textos escritos em Inglês. Dessa maneira, é necessário estimar o sentimento associado para cada menção às entidades nos textos publicados em idioma Português do Valor Econômico. Essa estimativa é obtida através da média dos sentimentos das sentenças onde as menções aparecem, ponderando a influência das respectivas sentenças para o sentimento geral de cada documento.

A atividade *Calcular Indicadores e Métricas* calcula as métricas de desempenho, tanto para entidades individualmente como para grupos de entidades semanticamente relacionadas (um componente conexo para cada um dos seis partidos considerados, com seus respectivos filiados políticos). Ambos os estudos de casos são submetidos, inicialmente, as métricas de desempenho consolidadas de f_1 a f_7, porém os resultados obtidos são semelhantes entre as métricas. Dessa maneira, o estudo de caso da intenção de voto (Seção 5.1.1) utiliza as métricas de desempenho consolidadas f_1 e f_2, pois são as métricas que mais se destacam. Já o estudo de caso da flutuação da moeda (Seção 5.1.2) utiliza a métrica f_3. Essas métricas são apresentadas no Quadro 3 do Capítulo 4.2. O algoritmo genético, responsável por ajustar os pesos das entidades, é executado com os seguintes parâmetros: os indivíduos (pesos) são

[4] https://www.bcb.gov.br/

representados por 7 dígitos; cada população possui 5000 indivíduos; o coeficiente de correlação de Pearson desejado deve ser superior a .75, acurácia de predição superior a .70; um limite de 400 gerações; e a taxa de mutação de .01, i.e., apenas 1% dos cruzamentos sofrem mutação.

A atividade *Correlacionar Métricas Calculadas com Medidas Reais* aplica o coeficiente de correlação de Pearson. Finalmente, a *Etapa 5 (Prediction)* utiliza o algoritmo OneR para ambos os estudos de caso, porém o estudo de caso da flutuação da moeda executa também Regressão Linear Simples. Ambos os algoritmos são implementados no Weka[5]. Esses algoritmos são executados em esquema de validação cruzada k-fold, com $k = 10$ (padrão Weka). Em outras palavras, o conjunto de dados de treinamento é dividido em 10 partes de mesmo tamanho. Para cada execução, uma parte é reservada para teste e as demais são utilizadas para treinamento. A média da acurácia das 10 execuções é exibida nos resultados (Capítulo 6).

5.2.3 Recursos de Hardware e Software

A solução necessária para capturar os dados, calcular as métricas de desempenho consolidadas e executar as tarefas do método *5-Ions* é implementada em Java na sua versão 8, usando o Ambiente de Desenvolvimento Integrado (*Integrated Development Environment* - IDE) Eclipse Neon 3. A sumarização é implementada em Python na sua versão 3.5 com a adição da biblioteca Sumy. Os dados são armazenados no banco de dados não relacional MongoDB em sua versão 4. Todos os experimentos são executados em um computador pessoal equipado com um processador Intel Core i5 - 4300M 2 CPU 2.60GHz, disco rígido com 500GB, 8GB RAM, executando o sistema operacional Microsoft Windows de 64-bit versão 7 Profissional Service Pack 1.

[5] https://www.cs.waikato.ac.nz/ml/weka/

6 RESULTADOS E DISCUSSÃO

Este capítulo apresenta e discute os resultados dos experimentos dos estudos de caso de Intenção de Voto e da Flutuação da Moeda, descritos no capítulo anterior. A Seção 6.1 analisa o impacto da sumarização nos textos através do estudo de caso da Intenção de Voto, além disso, esse estudo é utilizado como linha de base para validação do método *5-Ions*. A Seção 6.2, por sua vez, exibe os resultados do estudo de caso da Flutuação da Moeda, sem considerar sumarização de texto, porém aplicando aprendizado de máquina e modelo estatístico para predição.

6.1 INTENÇÃO DE VOTO

O resultado do estudo de caso da Intenção de Voto verificou a influência da aplicação de sumarização nas correlações e predições. Essa seção está segmentada em *Textos Não Sumarizados*, i.e., a não execução da Etapa 1 do método *5-Ions* (Capítulo 4), *Textos Sumarizados com LSA* e *Análise Comparativa entre os Tempos de Execução*. Cada segmento, por sua vez, é compreendido por uma seção para retratar os resultados das correlações e outra seção para expor os resultados das predições, através da métrica de acurácia. A análise comparativa considera os tempos necessários para executar as atividades de sumarizar textos, NER & NED, analisar sentimento e analisar sentimento no contexto de entidade.

Cada período de tempo ρ e $\rho+1$ de aplicação da pesquisa é considerado no momento do cálculo da métrica consolidada, i.e., os textos analisados estão contidos nesse período de tempo. O ajuste automático dos pesos ocorre apenas para os partidos políticos. O processo de predição classificou a oscilação (aumenta, diminui ou permanece estável) da intenção de voto nos partidos políticos. Desta forma, o conjunto de dados de treinamento é particionado por instituto de pesquisa.

6.1.1 Textos Não Sumarizados

Os textos processados não sofreram nenhum tipo de sumarização nem pré-processamento. Dessa maneira, os textos estão escritos no idioma Inglês, resultado da tradução automática realizada com o Google Translate a partir do idioma Alemão.

6.1.1.1 Análise de Correlação

As Figuras 5 e 6 apresentam o coeficiente de correlação de Pearson das métricas de desempenho consolidadas f_1 e f_2 do método *5-Ions* para os dados de todas as 55 publicações dos institutos de pesquisa *Forschungsgruppe Wahlen* e *Emnid*, respectivamente. O eixo vertical representa o módulo dos coeficientes de correlação. Foi aferido a correlação da medida real de desempenho, obtida através dos institutos

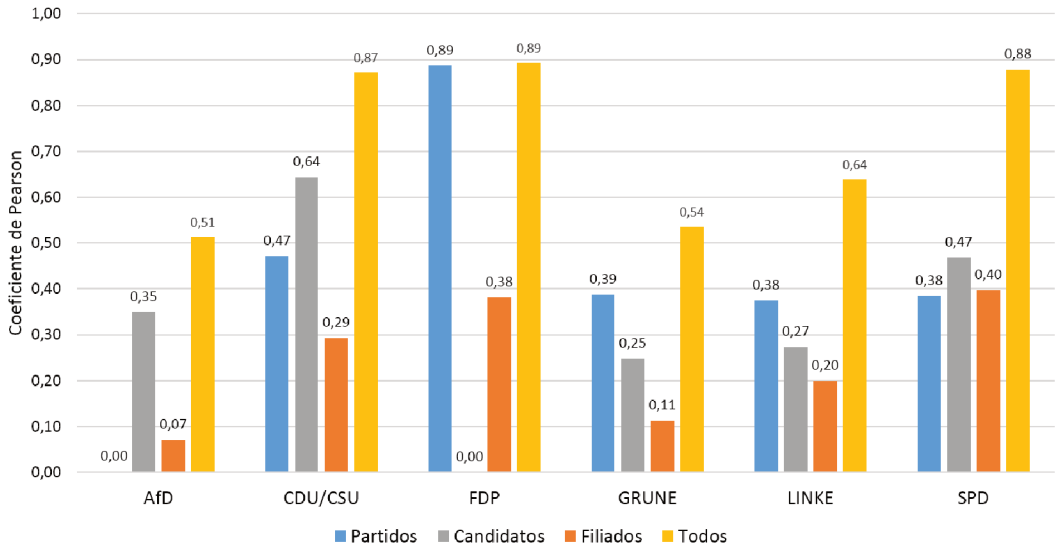
de pesquisa, com as métricas calculadas a partir de conjuntos de entidades distintas encontradas nos documentos das notícias. O cálculo das métricas considera os seguintes grupos de entidades: somente menções aos partidos (*Partidos*); somente menções aos candidatos à presidência (*Candidatos*); menções aos políticos filiados a cada partido (*Filiados*); e menções a todas as entidades de cada componente conexo (*Todos*). O partido AfD registrou valores de coeficiente próximos a zero para o instituto *Forschungsgruppe Wahlen*. Contudo, o candidato do partido FDP, Christian Lindner, não teve suas métricas de desempenho calculadas, pois foram identificados somente três menções ao candidato no período das análises.

A Figura 5a apresenta as correlações das pesquisas do instituto *Forschungsgruppe Wahlen* para a métrica f_1. Note que *Todos* obteve coeficientes de correlação superior em todos os partidos, iniciando sempre com valores maiores que .5, e se aproximando de .9 em ao menos três partidos. A Figura 5b apresenta as correlações das pesquisas do instituto *Forschungsgruppe Wahlen* para a métrica f_2. Seu comportamento é similar ao da métrica f_1, com correlações para *Todos* melhores que os demais, iniciando sempre acima de .45 porém abaixo de .9.

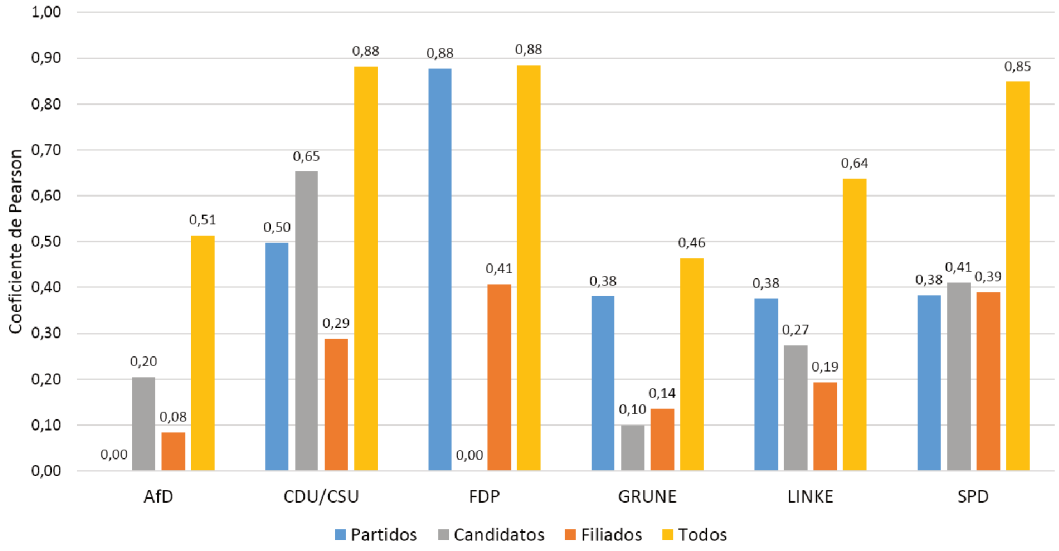
O instituto de pesquisa *Emnid* exibe comportamento análogo ao instituto *Forschungsgruppe Wahlen*. Esse comportamento pode ser percebido na comparação das Figuras 5 e 6. A métrica f_1, demonstrada na Figura 6a, contém coeficientes de correlação similares a métrica f_2, exposto na Figura 6b, exceto pelo partido AfD. Note que *Todos* possui os maiores valores de correlação dentre todos os partidos, iniciando sempre com valores superiores a .6, exceto para o partido AfD, e próximo a .8 para, ao menos, três partidos. O partido AfD apresenta correlações maiores que .5 para a métrica f_1 e conjunto de entidades *Todos*, e valores próximos de .4 para a métrica f_2.

As seguintes conclusões podem ser desenhadas para esses resultados: (i) as métricas consolidadas baseadas nas menções as entidades alvo e suas entidades associadas podem apresentar maiores correlações para medidas reais de desempenho do que métricas calculadas apenas para as menções a entidade alvo; (ii) levar em conta apenas as menções a entidade alvo pode não levar a uma correlação significativa entre sua métrica de desempenho e suas medidas reais de desempenho; (iii) levar em conta somente menções a entidades semanticamente relacionadas (*Filiados*), mas não as menções a entidade alvo, pode não apresentar correlações também; (iv) as entidades alternativas (e.g. candidatos vs. partidos), cuja as métricas de desempenho individual (i.e. calculada apenas das menções diretas da entidade) tem a correlação maior com as medidas reais de desempenho correspondentes, sugerem que métricas de desempenho consolidadas para entidades semanticamente relacionadas (do mesmo componente conexo na RS) pode beneficiar usuários, pois evita a difícil tarefa de escolher uma entidade alvo.

Figura 5 – Coeficiente de Correlação de Pearson - *Forschungsgruppe Wahlen*: (a) Métrica f_1 (b) Métrica f_2.



(a)



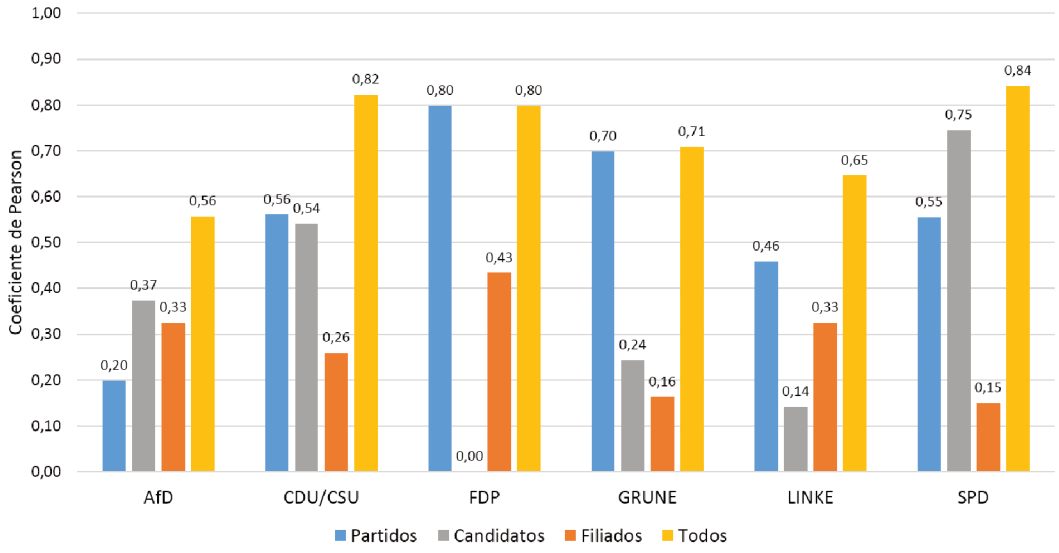
(b)

Fonte: elaborado pelo autor.

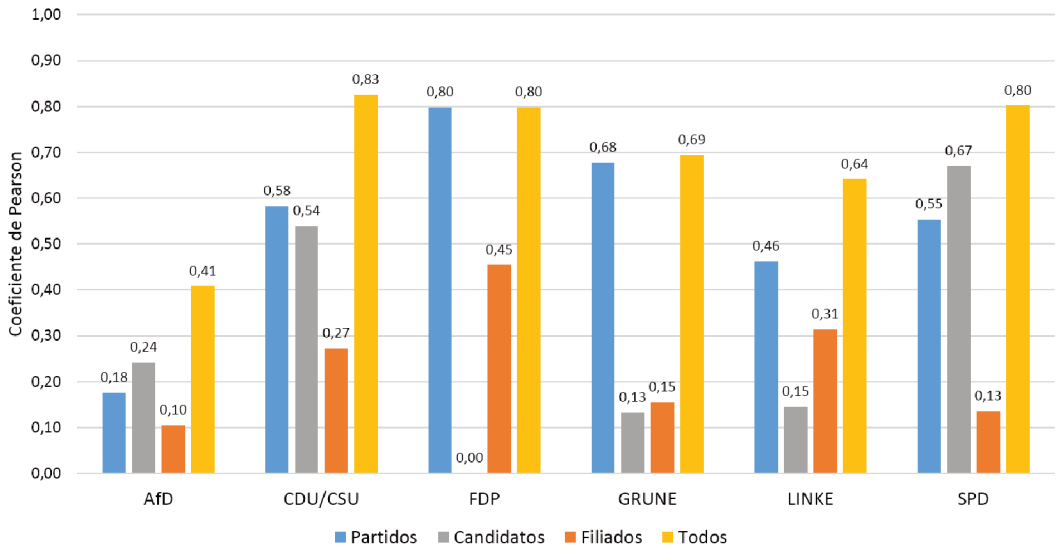
6.1.1.2 Análise Preditiva

As Figuras 7 e 8 apresentam as acurácias das predições obtidas através do uso das métricas de desempenho consolidadas f_1 e f_2 do método *5-Ions* para prever as mudanças do desempenho real fornecida pelas 55 pesquisas dos institutos *Forschungsgruppe Wahlen* e *Emnid*, respectivamente. O eixo vertical se refere aos valores das acurácias das predições. O eixo horizontal se refere aos três classificadores baseados no algoritmo OneR usado para prever as mudanças nas intenções de votos nos partidos políticos. O primeiro classificador, chamado *3 Classes (Aumenta, Dimi-*

Figura 6 – Coeficiente de Correlação de Pearson - Emnid: (a) Métrica f_1 (b) Métrica f_2.



(a)



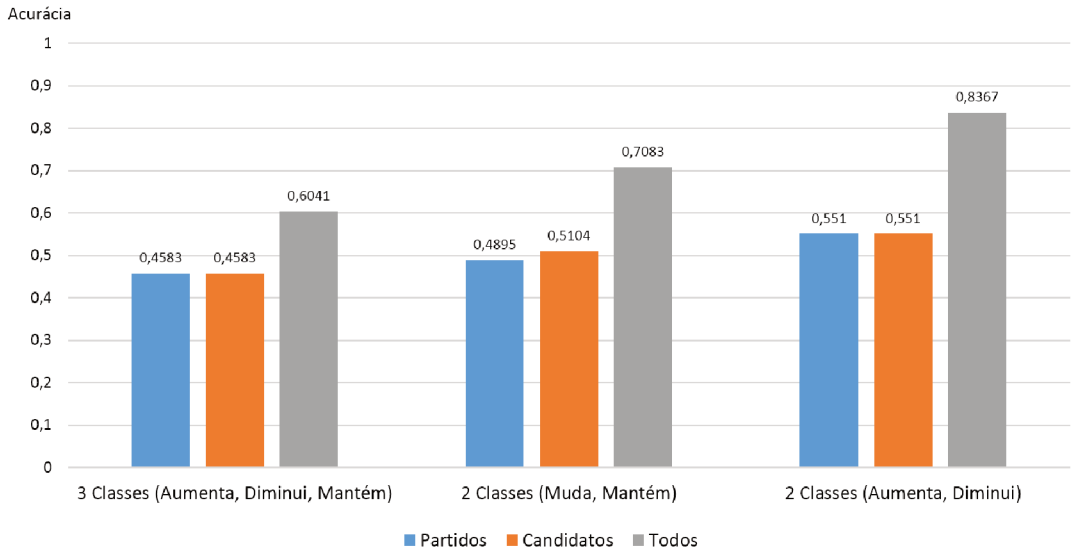
(b)

Fonte: elaborado pelo autor.

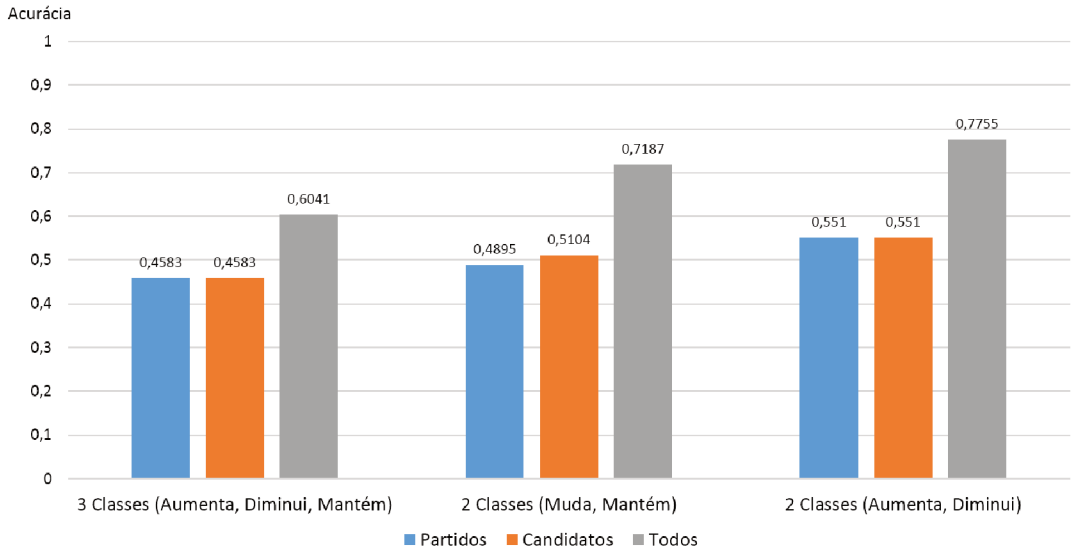
nui, Mantém), prevê se a intenção de voto em cada partido irá aumentar, diminuir ou permanecer estável. O segundo classificador, *2 Classes (Muda, Mantém)*, prevê se as intenções de votos irão mudar ou não. Finalmente, o terceiro classificador, *2 Classes (Aumenta, Diminui)*, prevê se a intenção de voto irá aumentar ou diminuir. O conjunto de treinamento para o classificador *2 Classes (Aumenta, Diminui)* não inclui dados por período em que não ocorre mudança no desempenho real do partido.

As medidas de desempenho usadas para fazer as predições estavam consolidadas para cada componente conexo na rede semântica (RS). Conjuntos alternativos de menções foram considerados para calcular as medidas: somente menções aos

Figura 7 – Acurácia das Predições - *Forschungsgruppe Wahlen*: (a) Métrica f_1 (b) Métrica f_2.



(a)



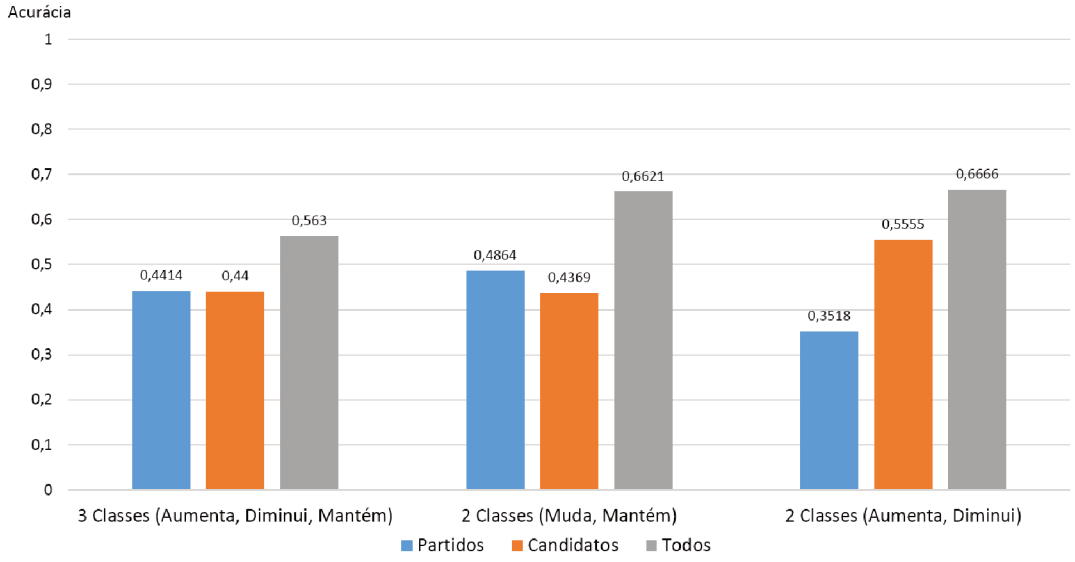
(b)

Fonte: elaborado pelo autor.

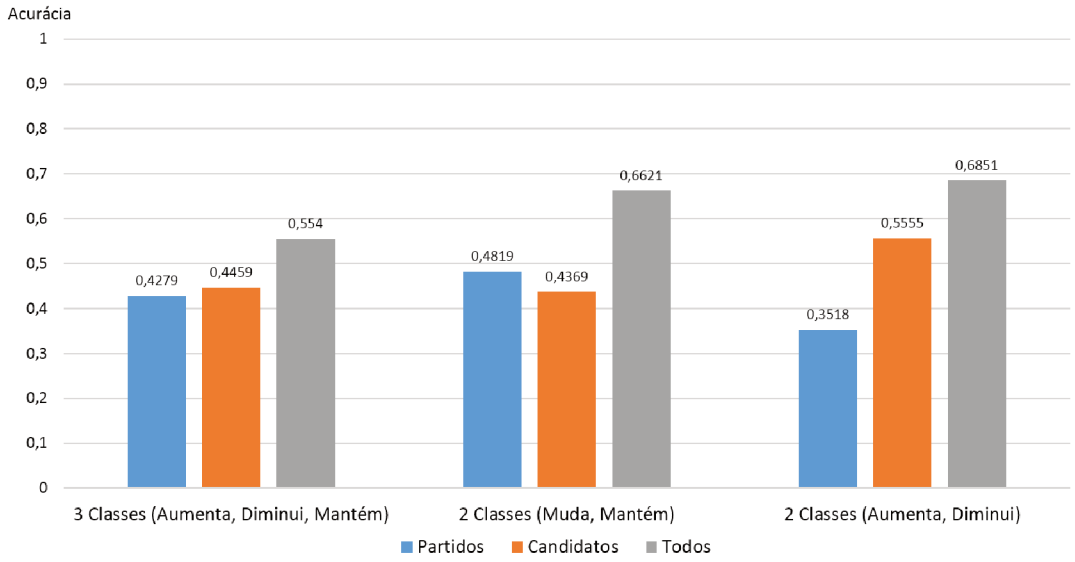
Partidos; somente menções aos *Candidatos*; e menções a todas as entidades do componente conexo que contém o partido político (*Todos*). O ajuste automático dos pesos das entidades ocorre somente para *Todos* e de forma independente para cada classificador.

A acurácia das predições para o instituto de pesquisa *Forschungsgruppe Wahlen* obtidas com as métricas de desempenho f_1 e f_2 são apresentadas nas Figuras 7a e 7b, respectivamente. As acurácias obtidas com essas duas métricas são muito similares. A métrica f_1 fornece predição ligeiramente melhor, especialmente quando considerado

Figura 8 – Acurácia das Predições - Emnid: (a) Métrica f_1 (b) Métrica f_2.



(a)



(b)

Fonte: elaborado pelo autor.

menções a *Todos* com o classificador *2 Classes (Muda, Mantém)*. Para ambas as métricas, é possível verificar a acurácia superior para menções a *Todos*. Considerando os resultados somente para a métrica f_1. A acurácia do classificador *3 Classes (Aumenta, Diminui, Mantém)* para menções a *Todos* é 41% maior que a acurácia de *Partidos*, e 65% maior que a acurácia de *Candidatos*. Situações similares ocorrem para os classificadores *2 Classes (Muda, Mantém)* e *2 Classes (Aumenta, Diminui)*, para qual *Todos* fornecem ganhos, respectivamente, de 54% e 55% quando comparados a *Partidos*, e 15% e 67% quando comparados a *Candidatos*.

A acurácia das predições para *Emnid GmbH & Co. KG (Emnid)* com as métricas

f_1 e f_2 são mostradas nas Figuras 8a e 8b, respectivamente. Para esse instituto de pesquisa, as acurácias obtidas usando a métrica f_2 são ligeiramente melhores em comparação os resultados da métrica f_1, principalmente quando se considera menções a *Todos* com o classificador *2 Classes (Muda, Mantém)*. Entretanto, a superioridade das menções a *Todos* permanece bem afiada. Considerando os resultados para a métrica f_1. A Figura 8a aponta que as acurácias do classificador *3 Classes (Aumenta, Diminui, Mantém)* são 27% maiores para menções a *Todos* quando comparadas tanto aos *Partidos* quanto aos *Candidatos*. Fazendo a mesma comparação para os classificadores *2 classes (Muda, Mantém)* e *2 classes (Aumenta, Diminui)*, observa-se que os ganhos das menções a *Todos* são, respectivamente, de 36% e 89% quando comparados aos *Partidos*, e de 51% e 20% quando comparados aos *Candidatos*.

Já a Figura 8b mostra a métrica f_2. As acurácias aferidas para *Todos* é superior independente do classificador. No classificador *3 Classes (Aumenta, Diminui, Mantém)* os ganhos são cerca de 29%, quando comparado *Todos* com *Partidos* e próximas a 24% para *Candidatos*. Comparação análoga para os classificadores *2 Classes (Muda, Mantém)* e *2 Classes (Aumenta, Diminui)*, respectivamente, revela ganhos nas acurácias em torno de 37% e 94% para os *Partidos*, e próximas a 51% e 23% para os *Candidatos*.

Para certas métricas, classificadores e institutos de pesquisa, considerar somente menções aos *Candidatos* produz melhores resultados do que considerar somente menções aos *Partidos*. Entretanto, isso não ocorre com todos os classificadores. De outra maneira, considerando todas as menções relacionadas as entidades do tipo partido (i.e., políticos no componente conexo da RS referindo-se ao respectivo partido político) sempre fornece melhores valores de acurácia.

Em todos os casos testados, o classificador *3 Classes (Aumenta, Diminui, Mantém)* mostrou os piores valores de acurácia, comparados com os outros classificadores. Contudo os classificadores *2 Classes (Muda, Mantém)* e *2 Classes (Aumenta, Diminui)*, que cobrem menos classes de predição do que o classificador *3 Classes (Aumenta, Diminui, Mantém)*, apresentam uma assertividade superior.

6.1.2 Textos Sumarizados com LSA

Aplicando a *Etapa 1 (Summarization)* do método *5-Ions*, que é opcional, pôde-se verificar a influência de sumarização na estimativa da intenção de votos. A técnica de sumarização aplicada foi a Análise Semântica Latente (*Latent Semantic Analysis* - LSA). Os textos foram reduzidos a 30% do seu tamanho original. As demais condições do experimento são idênticas ao da seção anterior.

6.1.2.1 Análise de Correlação

As Figuras 9 e 10 exibem as correlações de Pearson entre as métricas consolidadas f_1 e f_2 e as pesquisas de opinião publicadas pelos institutos *Forschungsgruppe Wahlen* e *Emnid*, respectivamente. O eixo vertical representa o módulo dos coeficientes de correlação. Os conjuntos de entidades consideradas nas análises são: somente menções aos partidos (*Partidos*); somente menções aos candidatos à presidência (*Candidatos*); menções aos políticos filiados a cada partido (*Filiados*); e menções a todas as entidades de cada componente conexo (*Todos*). As entidades filiadas ao partido AfD e ao partido LINKE não obtiveram menções identificadas nos textos sumarizados, o mesmo aconteceu para o partido FDP e seu candidato Christian Lindner, por isso seus coeficientes de correlação foram considerados zero.

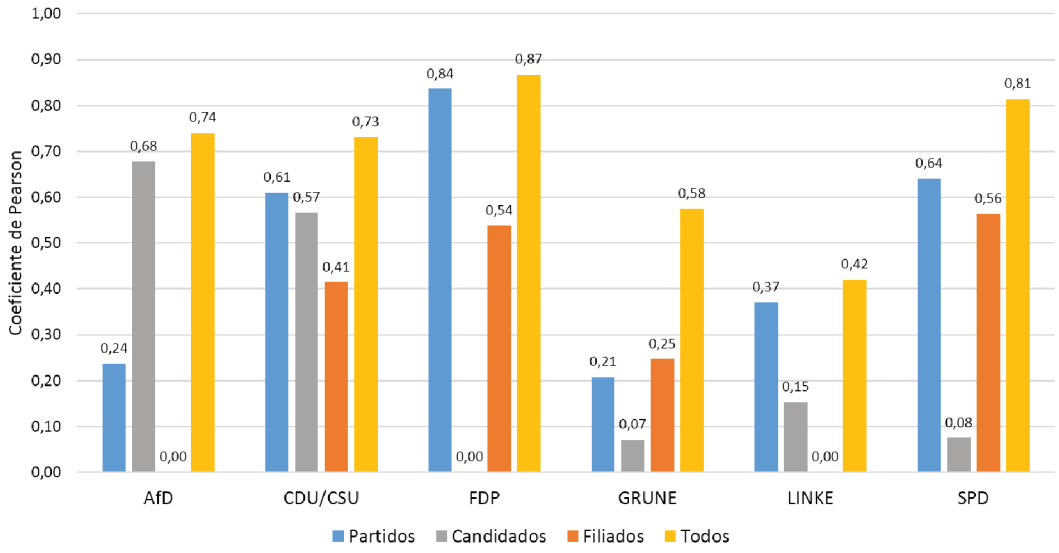
A Figura 9a apresenta as correlações entre as pesquisas divulgas pelo instituto *Forschungsgruppe Wahlen* e a métrica consolidada f_1. Note que *Todos* sempre apresenta os melhores valores de correlação com valores superior a .70 em quatro dos seis partidos considerados. O partido FDP obtém uma alta correlação para menções ao *Partido* (.84), porém menções a *Todos* consegue ter uma correlação ainda melhor, .87. Já o partido LINKE registrou o pior coeficiente de correlação para *Todos* (.42). A Figura 9b, por sua vez, exibe as correlações das pesquisas do instituto *Forschungsgruppe Wahlen* para a métrica f_2. As correlações para *Todos* mantém comportamento similar ao da métrica f_1, com oscilações de até 3% entre as menções correspondentes, para mais ou para menos, exceto para o partido LINKE, onde menções aos *Candidatos* supera menções aos *Partidos*.

O instituto de pesquisa *Emnid* expõe semelhanças entre as correlações obtidas com a métrica consolidada f_1 e f_2, Figuras 10a e 10b, respectivamente. Os coeficientes medidos por entidades apresentaram pequenas oscilações, ou mesmo nenhuma oscilação, entre os gráficos. De maneira geral, as correlações para *Todos* obtiveram os melhores valores, quando comparado com as demais entidades, variando entre .44 e .82, exceto para o partido GRUNE que registrou valores inferior a .3.

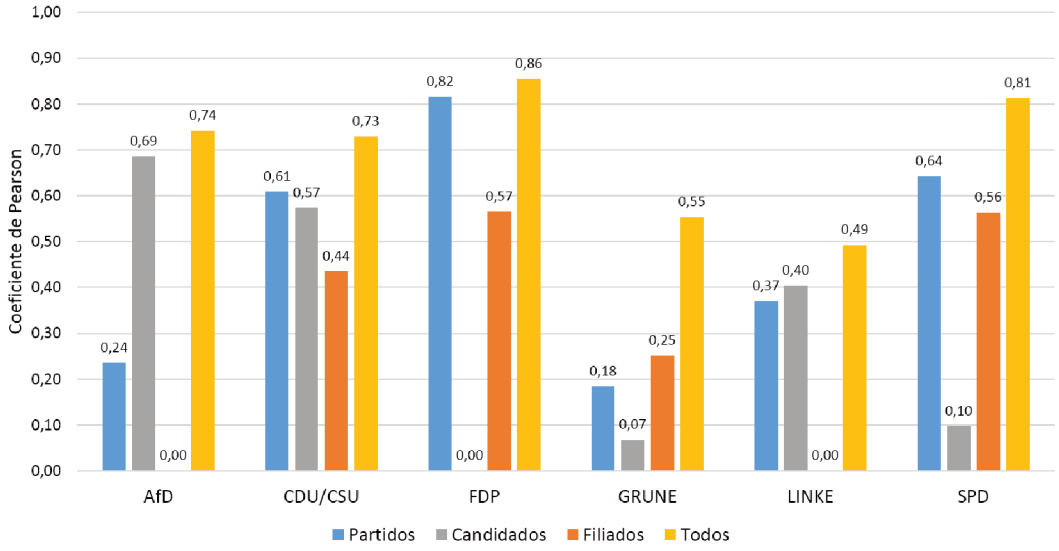
Quando comparado os resultados obtidos pelos textos originais (sem sumarização) e os textos sumarizados com LSA, com as mesmas métricas consolidadas, percebe-se ganhos e perdas. Entidades com poucas menções no textos originais torna-se um problema para a sumarização. Esse problema foi percebido nas menções ao partido AfD e nas menções aos filiados dos partidos AfD e LINKE.

Comparando somente a métrica consolidada f_1 verifica-se ganhos na utilização de sumarização em 7 das 23 medições (candidatos do partido FDP não obtiveram menções identificadas em nenhuma dos conjuntos de dados). Já para as menções aos *Partidos*, os textos sumarizados obtêm melhores resultados em somente 2 partidos e perdas em 4 partidos, os ganhos alcançam 70% para o partido AfD e as perdas atingem até 94% para o partido GRUNE. Verificando as menções aos *Filiados*, os

Figura 9 – Coeficiente de Correlação de Pearson para os Textos Sumarizados - *Forschungsgruppe Wahlen*: (a) Métrica f_1 (b) Métrica f_2.



(a)

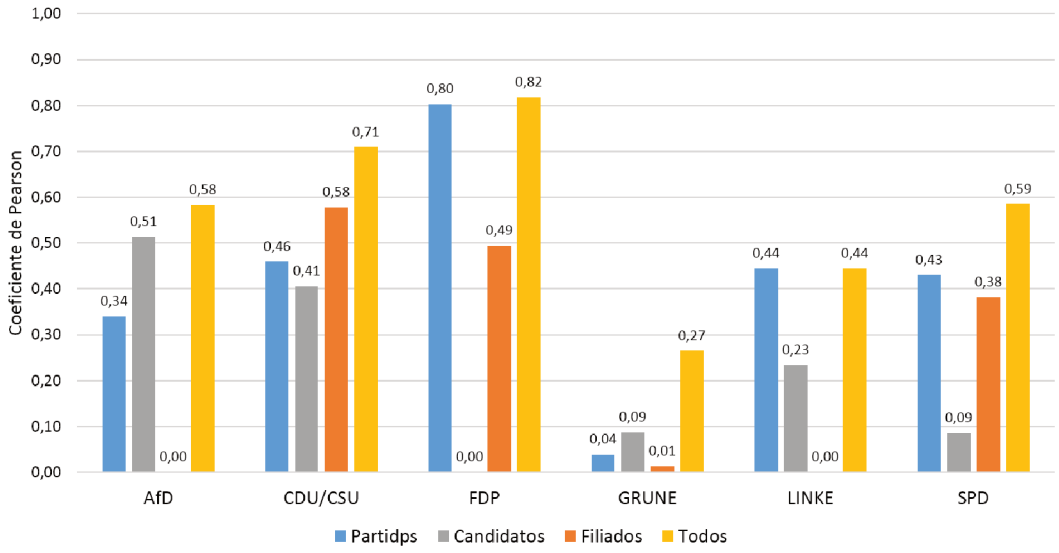


(b)

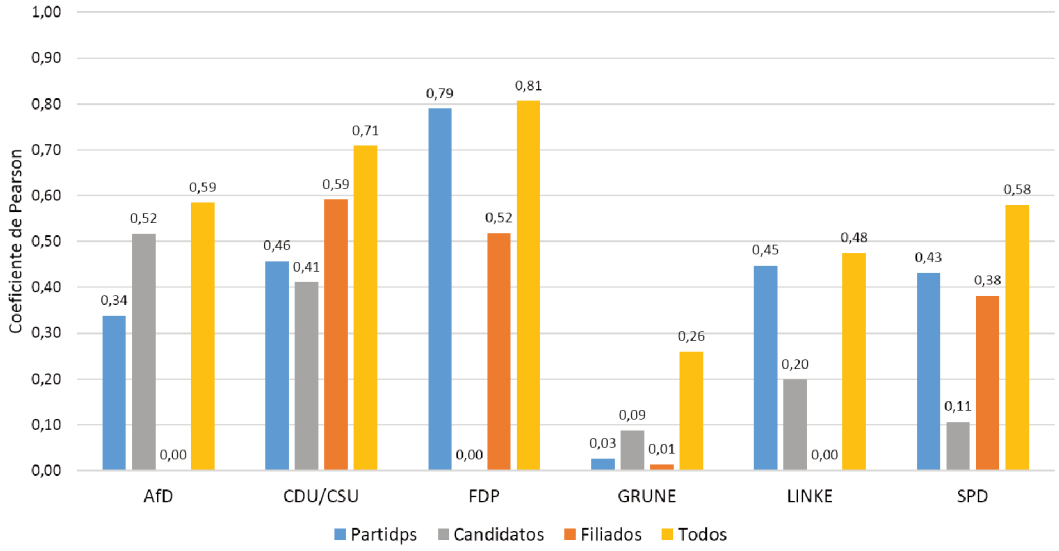
Fonte: elaborado pelo autor.

textos sumarizados obtêm melhores resultados em 3 partidos e piora em outros 3 partidos, os ganhos atingem 153% para o partido SPD e a perda alcança 100% nos partidos AfD e LINKE. Em relação aos *Candidatos*, os textos sumarizados registram melhores resultados em 2 partidos e piora em outros 3 partidos, com ganho máximo de 64% para o partido LINKE e perda de até 89% para o partido SPD. Por fim, em relação as menções a *Todas* as entidades, os textos sumarizados obtêm melhores resultados em 2 partidos e piora nos demais, o maior ganho atinge somente 4% para o partido AfD em contrapartida a perda alcança 62% para o partido GRUNE.

Figura 10 – Coeficiente de Correlação de Pearson para os Textos Sumarizados - Emnid: (a) Métrica f_1 (b) Métrica f_2.



(a)



(b)

Fonte: elaborado pelo autor.

Realizando a mesma comparação para a métrica consolidada f_2 percebe-se ganhos na utilização de sumarização somente em 8 das 23 medições. Ao analisar menções aos *Partidos*, os textos sumarizados obtêm melhor resultado apenas para um partido e perda nos demais, o ganho acontece no partido AfD (92%) e a perda alcança 96% para o partido GRUNE. Já para as menções aos *Filiados*, os textos sumarizados obtêm melhores resultados em 3 partidos e piora em outros 3 partidos, o ganho máximo atinge 182% para o partido SPD e a perda alcança 100% nos partidos AfD e LINKE. Em relação as menções aos *Candidatos*, os textos sumarizados registram

melhores resultados em 2 partidos e piora em outros 3, o ganho máximo é de 113% para o partido AfD e a perda atinge 84% para o partido SPD. Por fim, as menções a *Todas* as entidades, os textos sumarizados obtêm melhores resultados em 2 partidos e perda nos demais, os ganhos alcançam 43% para o partido AfD e a perda atinge 62% para o partido GRUNE.

Os partidos GRUNE e LINKE figuram entre as maiores perdas devido à baixa presença de menções as suas entidades nos textos analisados, independente da métrica consolidada aplicada e do uso ou não de sumarização de textos. A inconstância entre perdas e ganhos ao se aplicar sumarização necessita de maiores experimentos para ter sua causa identificada. O uso de técnicas mais sofisticadas de sumarização de texto poderá obter melhores resultados nas correlações, porém a comparação de técnicas não é abordada no escopo deste trabalho. De toda forma, a comparação das correlações dá indícios de inviabilidade do uso de sumarização, ao menos para esse estudo de caso.

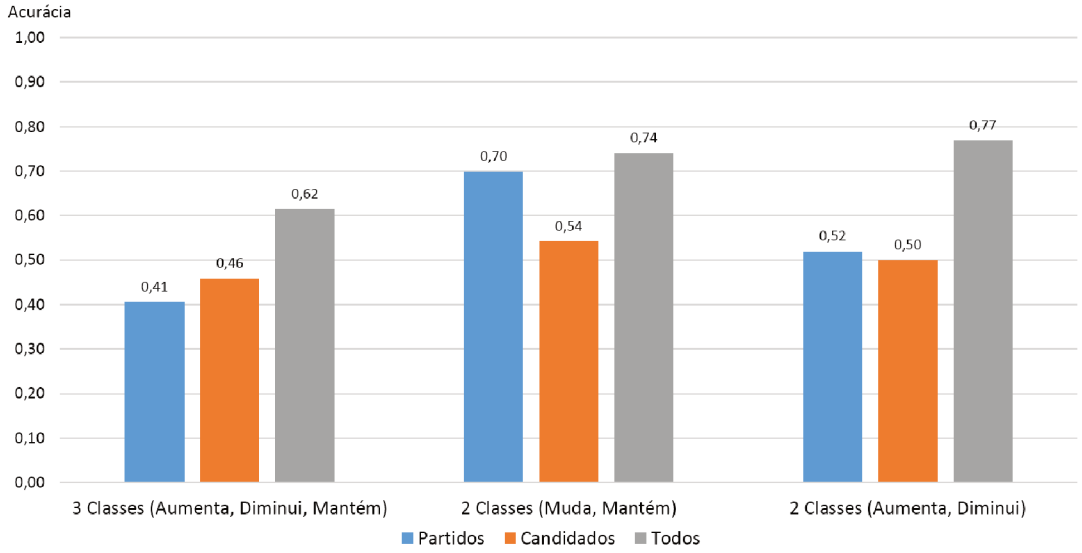
6.1.2.2 Análise Preditiva

As Figuras 11 e 12 apresentam as acurácias das predições obtidas através das métricas consolidadas f_1 e f_2 do método *5-Ions* e das publicações das intenções de voto dos institutos de pesquisa *Forschungsgruppe Wahlen* e *Emnid*, respectivamente. O eixo vertical se refere aos valores das acurácias das predições. O eixo horizontal se refere aos três classificadores baseados no algoritmo OneR usado para prever as mudanças nas intenções de votos nos partidos políticos. O primeiro classificador, chamado *3 Classes (Aumenta, Diminui, Mantém)*, prevê se a intenção de voto em cada partido irá aumentar, diminuir ou permanecer estável. O segundo classificador, *2 Classes (Muda, Mantém)*, prevê se as intenções de votos irão mudar ou não. Finalmente, o terceiro classificador, *2 Classes (Aumenta, Diminui)*, prevê se a intenção de voto irá aumentar ou diminuir. O conjunto de treinamento para o classificador *2 Classes (Aumenta, Diminui)* não inclui dados por período em que não ocorre mudança no desempenho real do partido.

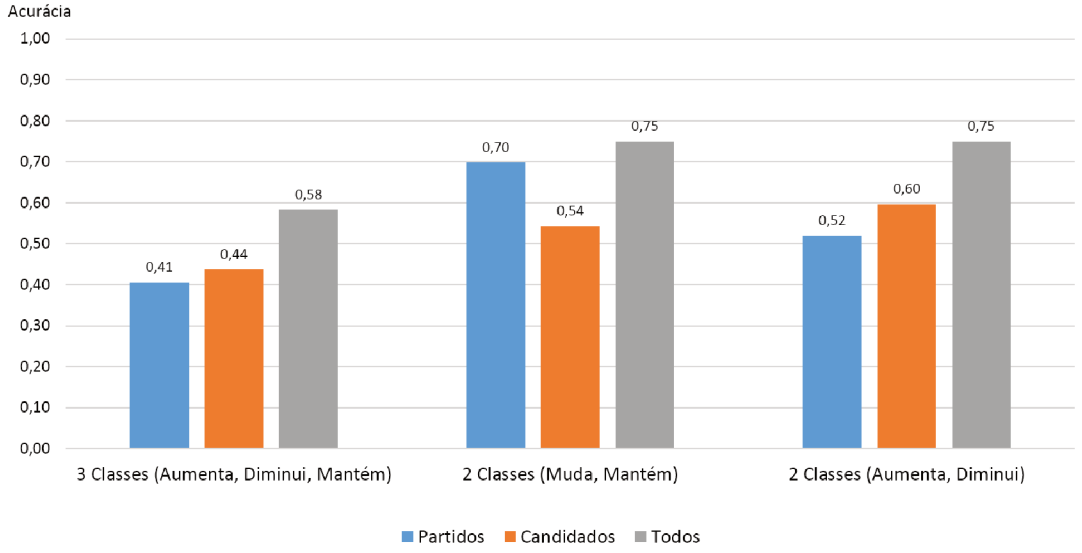
As medidas de desempenho utilizadas para fazer as predições são consolidadas para cada componente conexo da rede semântica (RS). Os conjuntos de menções considerados no cálculo das medidas são: somente menções aos *Partidos*; somente menções aos *Candidatos*; e menções a todas as entidades do componente conexo que contém o partido político (*Todos*). O ajusto automático dos pesos ocorre de forma independente por classificado e apenas para o conjunto que contém todas as entidades por partido político (*Todos*).

A acurácia das predições para o instituto de pesquisa *Forschungsgruppe Wahlen* obtidas com as métricas de desempenho f_1 e f_2 são apresentadas nas Figuras 11a e 11b, respectivamente. As acurácias obtidas com essas duas métricas são

Figura 11 – Acurácia das Predições para os Textos Sumarizados - *Forschungsgruppe Wahlen*: (a) Métrica f_1 (b) Métrica f_2.



(a)

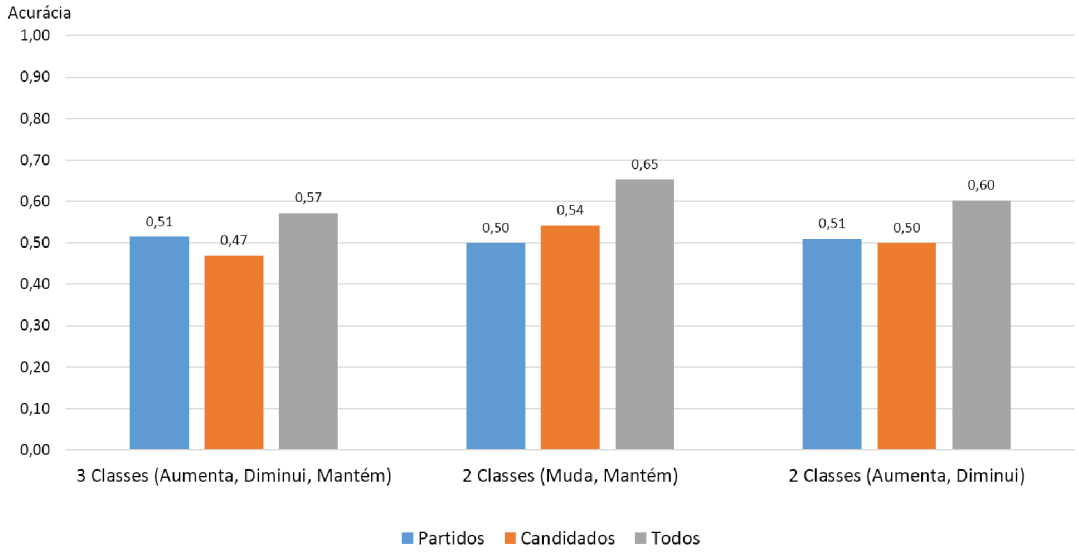


(b)

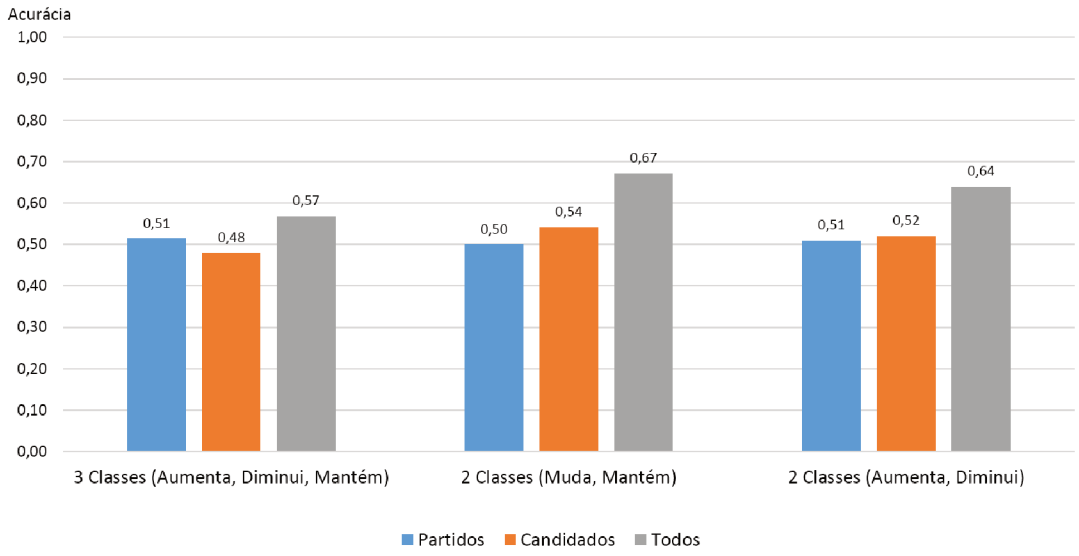
Fonte: elaborado pelo autor.

similares. A métrica f_1 fornece predição ligeiramente melhor, especialmente quando considerado menções a *Todos* com o classificador *2 Classes (Aumenta, Diminui)*. Para ambas as métricas, é possível verificar acurácia superior para menções a *Todos*. Considerando os resultados somente para a métrica f_1. A acurácia do classificador *3 Classes (Aumenta, Diminui, Mantém)* para menções a *Todos* é 51% maior que a acurácia de *Partidos*, e 34% maior que a acurácia de *Candidatos*. Situações similares ocorre para os classificadores *2 Classes (Muda, Mantém)* e *2 Classes (Aumenta, Diminui)*, para qual *Todos* fornecem ganhos, respectivamente, de 5% e 7% quando comparados a

Figura 12 – Acurácia das Predições para os Textos Sumarizados - Emnid: (a) Métrica f_1 (b) Métrica f_2.



(a)



(b)

Fonte: elaborado pelo autor.

Partidos, e 37% e 39% quando comparados a *Candidatos*.

A acurácia das predições para *Emnid GmbH & Co. KG (Emnid)* com as métricas f_1 e f_2 são mostradas nas Figuras 12a e 12b, respectivamente. Para esse instituto de pesquisa, as acurácias obtidas usando a métrica f_2 são ligeiramente melhores em comparação os resultados da métrica f_1. Entretanto, sempre há superioridade das menções a *Todos*. Considerando os resultados para a métrica f_1. A Figura 12a aponta que as acurácias do classificador *3 Classes (Aumenta, Diminui, Mantém)* é 12% maior para menções a *Todos* quando comparadas aos *Partidos* e 21% maior quando

comparado aos *Candidatos*. Fazendo a mesma comparação para os classificadores *2 classes (Muda, Mantém)* e *2 classes (Aumenta, Diminui)*, observa-se que os ganhos das menções a *Todos* são, respectivamente, de 30% e 18% quando comparados aos *Partidos*, e de 20%, para ambos, quando comparados aos *Candidatos*.

Já a Figura 12b também mostra que a métrica f_2 fornece acurácias superiores para *Todos*. *Todos* obtêm acurácias cerca de 12% superior para *Partidos* e próximas a 19% para *Candidatos*, usando o classificador *3 Classes (Aumenta, Diminui, Mantém)*. Comparação análoga para os classificadores *2 Classes (Muda, Mantém)* e *2 Classes (Aumenta, Diminui)*, respectivamente, revela acurácias em torno de 34% e 26% para os *Partidos*, e próximas a 24% e 23% para os *Candidatos*.

Quando comparado os resultados obtidos pelos textos originais (sem sumarização) e os textos sumarizados com LSA, e observando somente a métrica f_1, percebe-se ganhos de acurácia para textos sumarizados em dois classificadores testados e perdas somente em um classificador. Os textos sumarizados, no classificador *3 Classes (Aumenta, Diminui, Mantém)*, apresentaram ganhos de 11% e 1% em relação as menções aos *Partidos* e a *Todos*, respectivamente, e a mesma acurácia para as menções aos *Candidatos*. Já para o classificador *2 Classes (Muda, Mantém)* os textos sumarizados obtiveram ganhos de 42%, 6% e 4% para menções aos *Partidos*, *Candidatos* e a *Todos*, respectivamente. Por fim, para o classificador *2 Classes (Aumenta, Diminui)* os textos sumarizados registraram perdas de acurácia na predição de 5%, 9% e 8% para as menções aos *Partidos*, *Candidatos* e a *Todos*, respectivamente.

A métrica f_2 tem comportamento similar ao da métrica f_1, com ganhos de acurácia para textos sumarizados em dois classificadores e perdas somente para um classificador. Menções a *Partidos*, nos textos sumarizados, apresentam ganhos de 11% e 42% para os classificadores *3 Classes (Aumenta, Diminui, Mantém)* e *2 Classes (Muda, Mantém)*, respectivamente, e perda de 5% no classificador *2 Classes (Aumenta, Diminui)*. Já menções aos *Candidatos* registram ganhos de 4% e 6% para os classificadores *3 Classes (Aumenta, Diminui, Mantém)* e *2 Classes (Muda, Mantém)*, respectivamente, e perda de 3% no classificador *2 Classes (Aumenta, Diminui)*. Por fim, menções a *Todos* obtêm ganhos em todos os classificadores, os ganhos foram de 3%, 4% e 8% nos classificadores *3 Classes (Aumenta, Diminui, Mantém)*, *2 Classes (Muda, Mantém)* e *2 Classes (Aumenta, Diminui)*.

As grandes oscilações apresentadas nos coeficientes de correlações não são apresentadas nas acurácias dos classificadores. As perdas de acurácia para os textos sumarizados são percebidas somente para o classificador *2 Classes (Aumenta, Diminui)*, porém para a métrica f_2 ao se analisar menções a todas as entidades dos componentes conexos da RS (*Todos)* sempre se obtêm ganhos, alcançando até 8%. Os resultados obtidos para as predições com os textos sumarizados são satisfatórios, porém é necessário considerar o classificador utilizado para realizar a predição.

Ressalva-se que o ajuste da técnica de sumarização utilizada e de uma maior diversidade de fontes de dados podem apresentar melhores resultados.

6.1.3 Análise Comparativa entre os Tempos de Execução

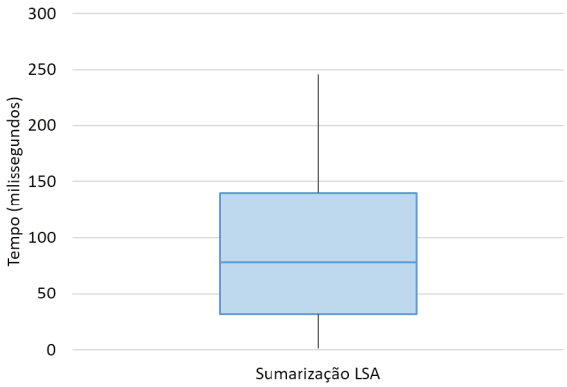
Esta seção compara os tempos de execução dos resultados obtidos com as análises dos textos originais, sem sumarização, e com as análises dos textos sumarizados com a técnica LSA. A Figura 13 apresenta os gráficos de caixa que correspondem aos tempos das execuções das tarefas: Sumarizar Textos, NER/NED, Analisar Sentimentos, Analisar Sentimentos no contexto de Entidade e Tempo Total para execução dessas tarefas. O eixo vertical corresponde ao tempo em milissegundos. O eixo horizontal exibe o conjunto de dados (sem sumarização e sumarização LSA) submetidos a execução, exceto a Figura 13a onde está representado o tempo de execução da sumarização. Os tempos foram submetidos ao teste U de Mann-Whitney para avaliar a existência de diferença significativa entre os valores medidos. Todos os tempos apresentaram diferenças significativas entre as amostras para uma significância de 5%. Todos os *outliers* foram removidos das análises dos tempos.

A Figura 13a apresenta os tempos necessários para executar a sumarização LSA nos textos. A mediana dos tempos foi de 46 milissegundos, oscilando entre 1 milissegundo e 106 milissegundos, tempo mínimo e tempo máximo, respectivamente. Os quartis inferior e superior foram de 31 e 62 milissegundos, respectivamente. Essas medidas permitem concluir a baixa interferência da sumarização no tempo total de processamento do texto, uma vez que 75% dos textos sumarizados tiveram tempo de execução de até 62 milissegundos.

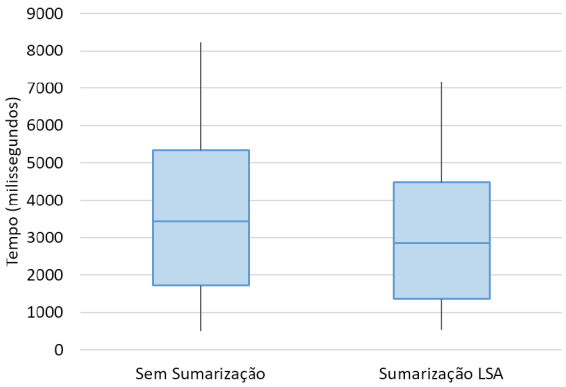
Já a Figura 13b apresenta os tempos necessários para realizar as anotações *NER/NED* para os documentos sem e com sumarização. Os tempos anotados para os documentos originais (i.e. sem sumarização) foram: tempo mínimo de 485 milissegundos, quartil inferior de 1235.5 milissegundos, mediana de 1712 milissegundos, quartil superior de 1902 milissegundos e tempo máximo de 2897 milissegundos. Já os tempos para realizar as mesmas atividades para os documentos sumarizados foram: tempo mínimo de 517 milissegundos, quartil inferior de 838 milissegundos, mediana de 1503 milissegundos, quartil superior de 1613 milissegundos e tempo máximo de 2705 milissegundos. Ao compararmos as medianas dos tempos percebemos uma diminuição de 12% nos tempos para realizar as anotações nos documentos sumarizados.

A Figura 13c exibe os tempos aferidos para realizar a atividade *Analisar Sentimentos*. Os tempos registrados para os documentos originais foram: tempo mínimo de 765 milissegundos, quartil inferior de 1765 milissegundos, mediana de 2122 milissegundos, quartil superior de 2431.5 milissegundos e tempo máximo de 3431 milissegundos. Já para os documentos sumarizados os tempos foram: tempo mínimo de 500 milissegundos, quartil inferior de 969 milissegundos, mediana de 1578 milissegundos, quartil

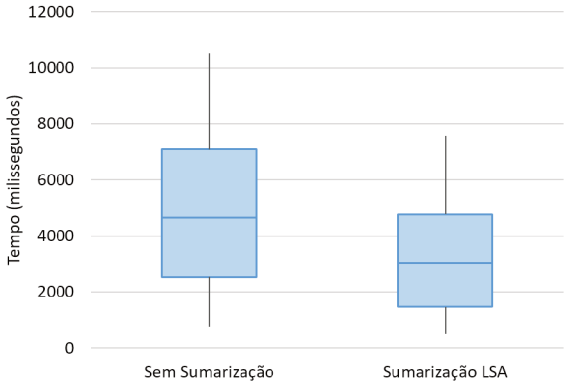
Figura 13 – Comparativo das Variações dos Tempos de Execução do Método *5-Ions* aplicando, ou não, Sumarização.



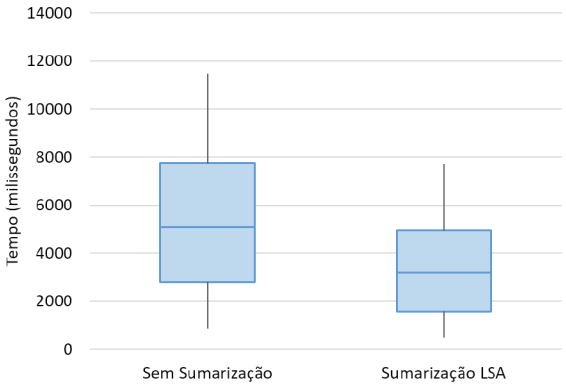
(a) Sumarização LSA



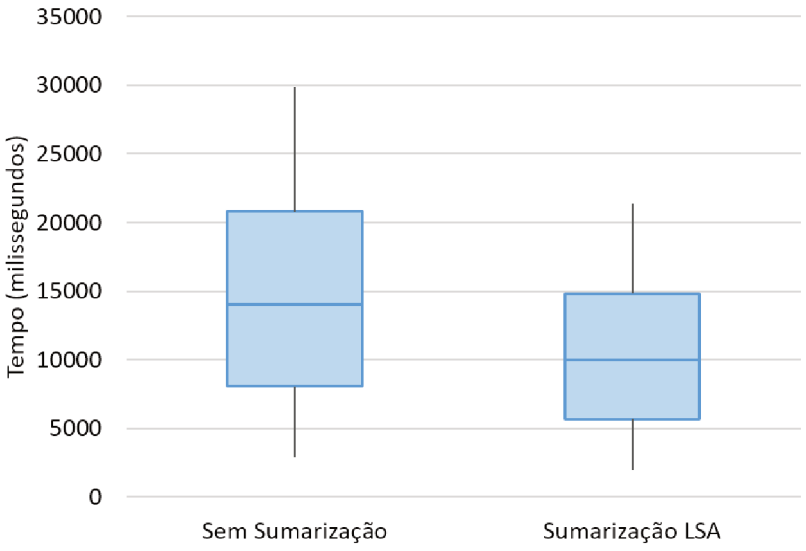
(b) *NER & NED*



(c) Análise de Sentimentos



(d) Análise de Sentimento no Contexto de Entidade



(e) Tempos Totais

Fonte: elaborado pelo autor.

superior de 1718 milissegundos e tempo máximo de 2797 milissegundos. A anotação

dos sentimentos para os textos sumarizados registraram menores tempos em todos os quartis, quando comparado aos textos originais. Comparando somente as medianas dos tempos podemos perceber uma diminuição de 25.6% entre os documentos sumarizados e os documentos originais.

As medições dos tempos percorridos para executar a análise de sentimentos no contexto de entidade são exibidas na Figura 13d. Os tempos anotados para os textos originais foram: tempo mínimo de 859 milissegundos, quartil inferior de 1937 milissegundos, mediana de 2297 milissegundos, quartil superior de 2656 milissegundos e tempo máximo de 3734 milissegundos. Já os tempos para realizar a mesma atividade, porém para os textos sumarizados, foram: tempo mínimo de 475 milissegundos, quartil inferior de 1094 milissegundos, mediana de 1609 milissegundos, quartil superior de 1766 milissegundos e tempo máximo de 2766 milissegundos. Comparando os tempos medidos entre os textos sumarizados e os textos originais temos reduções em todos os quartis. Analisando somente as medianas dos tempos verifica-se uma diminuição de 30% entre os textos sumarizados e os textos originais.

Por fim, o tempo total para a realização de todas as etapas aqui descritas é exibido na Figura 13e. Os textos originais registraram os seguintes tempos: tempo mínimo de 2856 milissegundos, quartil inferior de 5202 milissegundos, mediana de 5960 milissegundos, quartil superior de 6766 milissegundos e tempo máximo de 9095 milissegundos. Já os tempos para os textos sumarizados foram: tempo mínimo de 1979 milissegundos, quartil inferior de 3703 milissegundos, mediana de 4271 milissegundos, quartil superior de 4852 milissegundos e tempo máximo de 6573 milissegundos. Como já acontecia em todas as etapas, os tempos registrados para os textos sumarizados apresentaram melhores resultados em comparação aos textos originais, mesmo os textos sumarizados tendo executado a atividade adicional de sumarizar. Comparando somente a mediana dos tempos é percebido uma diminuição de 28% entre os textos sumarizados e os textos originais.

Os tempos aferidos nas atividades descritas nessa seção justificam a utilização de sumarização automática de texto. Os ganhos em tempo de execução chegam a 31% se compararmos os tempos totais mínimos, porém é necessário considerar também o impacto da sumarização nos coeficientes de correlação e nas acurácias das predições antes de concluir o uso, ou não, dessa técnica no método *5-Ions*. Os resultados da sumarização nas correlações e predições são exibidos nas próximas seções.

6.2 FLUTUAÇÃO DA MOEDA

A flutuação da moeda investiga oscilações na moeda Brasileira, o Real. Entretanto, são observados poucas menções diretas a entidade Real nos textos analisados, quando comparado ao número de menções as outras entidades que podem influenciar na sua cotação. Assim, esse experimento considera as menções a essas entidades,

classificadas de acordo com seus tipos (e.g., *Pessoa*, *Organização*) e papeis (e.g. uma Pessoa pode ter papel de legislador em alguma câmara ou de *Presidente* da república ou de alguma outra organização). Essa seção está dividida em *Análise de Correlação* e *Análise Preditiva*.

As análises, por sua vez, estão segmentadas em: entidades do tipo Pessoa; entidades do tipo Pessoa e com papel de Presidente; entidades do tipo Pessoa e com papel de Senador; entidades do tipo Pessoa e com papel de Ministro ou Diretor; entidades do tipo Organização; e todas as entidades. O período de tempo ρ e $\rho+1$ de cotação da moeda é considerado no momento do cálculo da métrica consolidada, i.e., os textos analisados estão contidos nesse período de tempo. A análise preditiva é realizada tanto para uma técnica de aprendizado de máquina quanto para um modelo estatístico. A técnica de aprendizado de máquina classifica as oscilações (aumenta, diminui ou permanece estável) da cotação da moeda. Já o modelo estatístico prevê os valores de cotação mediante as menções às entidades extraídas nos textos.

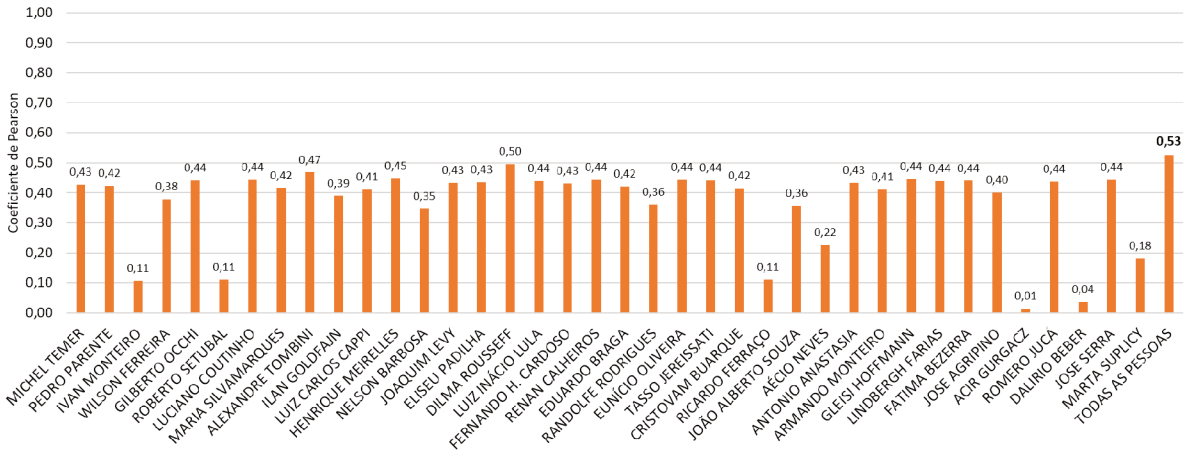
6.2.1 Análise de Correlação

Da Figura 14 até a Figura 18 são apresentadas as correlações de Pearson entre a métrica consolidada f_3, a partir dos textos publicados pelo jornal Valor Econômico, e a cotação da moeda Real disponibilizada pelo Banco Central Brasileiro (BACEN). O eixo horizontal de cada figura exibe as entidades analisadas. A última barra de cada um dos seis gráficos refere-se à correlação da cotação do Real com a métrica consolidada f_3 para todas as entidades consideradas nos respectivos gráficos. O gráfico da Figura 19 considera o coeficiente de correlação de Pearson entre a cotação do Real e cada entidade presente na RS considera, com a última barra representando todas essas entidades. A Figura 14 exibe o coeficiente de correlação para todas as entidades do tipo Pessoa. O gráfico da Figura 15 considera somente entidades do tipo Pessoa com papel de Presidente. A Figura 16 expõe somente entidades do tipo Pessoa com papel de Senador. A Figura 17 exibe o coeficiente de correlação das entidades do tipo Pessoa com papel de Ministro ou Diretor. Finalmente, a Figura 18 mostra as correlações entre a moeda Real e as entidades do tipo Organização.

Analisando somente menções as entidades do tipo Pessoa, Figura 14, é possível verificar a grande oscilação entre os coeficientes anotados. Alguns membros do legislativo com menor popularidade apresentaram menores coeficientes de correlação, por exemplo o senador Dalírio Beber de Santa Catarina com coeficiente de .04. Em contrapartida a então atual presidente da república Dilma Rousseff apresentou correlação de .50. Porém o maior coeficiente de correlação é registrado para a métrica consolidada com menções a todas as entidades, com coeficiente de .53.

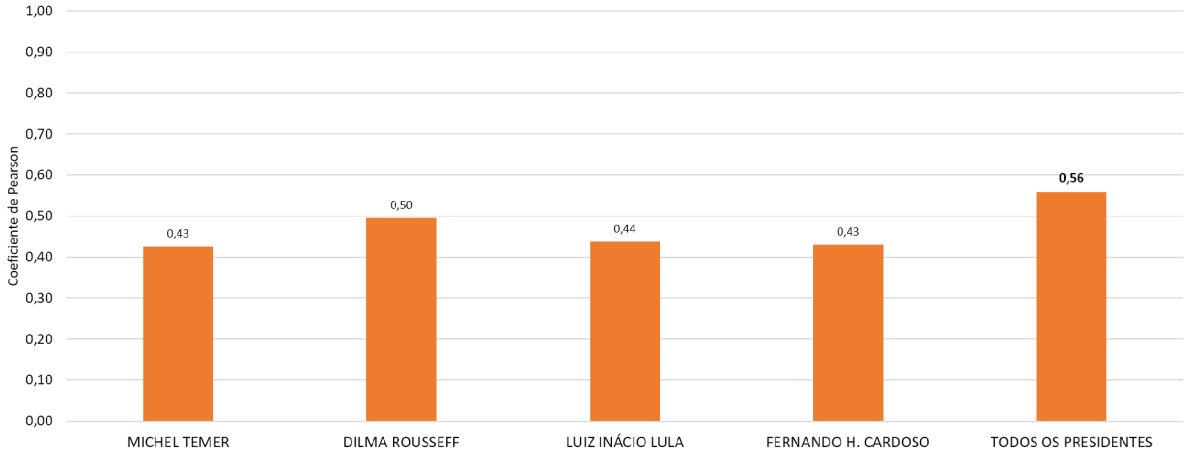
Já a Figura 15 exibe os coeficientes de correlação considerando somente menções as entidades do tipo Pessoa com papel de Presidente. Menções a entidade Dilma

Figura 14 – Coeficiente de Pearson somente para entidades do tipo Pessoa



Fonte: elaborado pelo autor.

Figura 15 – Coeficiente de Pearson somente para entidades do tipo Pessoa com papel de Presidente

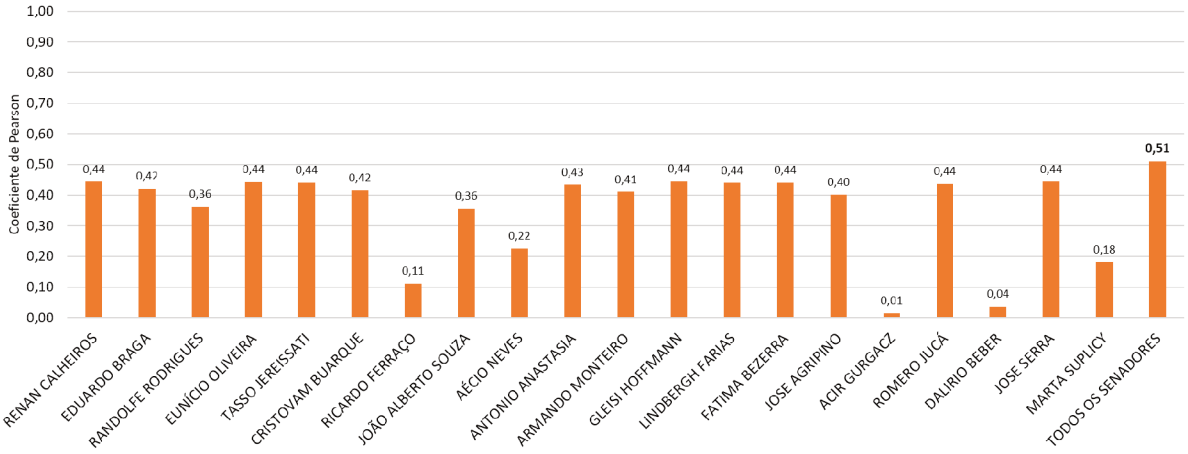


Fonte: elaborado pelo autor.

Rousseff, a presidente do Brasil durante o período analisado, que sofre um processo de *impeachment*, tem uma correlação pouco superior (.49) aos ex-presidentes considerados na análise (os demais possuem correlação em torno de .44). Entretanto, a correlação com a métrica consolidada é de .56, superior ao coeficiente obtidos ao se analisar todas as entidades do tipo Pessoa.

A Figura 16 apresenta os coeficientes de correlação entre a moeda Real e as entidades do tipo Pessoa com papel de Senador. Esse gráfico apresenta os menores coeficientes dessas análises pois conta com as entidades de menores popularidades, como já foi mencionado anteriormente. Percebe-se também coeficientes de correlação semelhantes entre os senadores que estão envolvidos em casos de corrupção (próximos a .44), que é o caso de Romero Jucá e José Serra, porém o senador Aécio

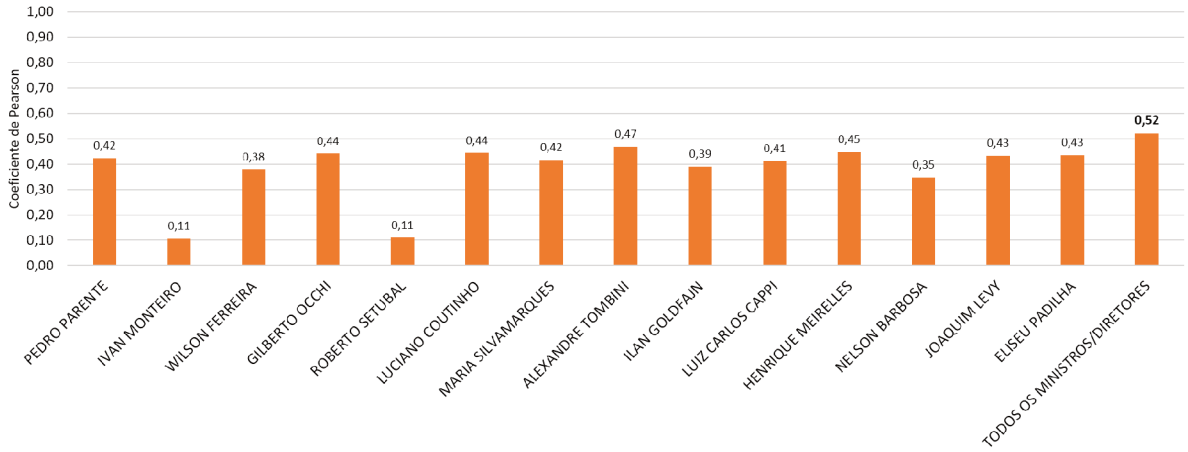
Figura 16 – Coeficiente de Pearson somente para entidades do tipo Pessoa com papel de Senador



Fonte: elaborado pelo autor.

Neves, que disputou a presidência na eleição anterior, possui um baixo coeficiente de correlação (.22). Já a correlação entre a moeda e a métrica calculada com base nas menções a todos os senadores obtém o melhor valor dentre todos anotados (.51).

Figura 17 – Coeficiente de Pearson somente para entidades do tipo Pessoa com papeis de Ministro ou Diretor

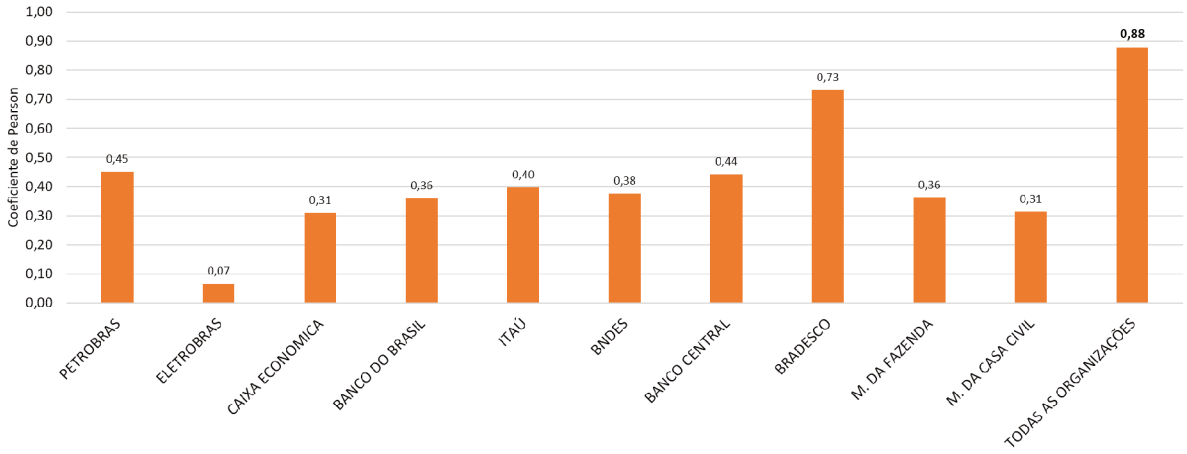


Fonte: elaborado pelo autor.

A Figura 17, por sua vez, exibe o gráfico com as análises de correlação entre a moeda Real e as menções aos principais ministros do governo e os principais diretores das empresas estatais ou instituições financeiras privadas. Tanto Ivan Monteiro, então presidente da Petrobras, quanto Roberto Setúbal, então presidente do banco Itaú, apresentaram coeficientes de correlações baixos, ambos próximos a .10. Já as demais entidades apresentaram coeficientes próximos entre si, com destaque para Alexandre Tombini, então presidente do Banco Central do Brasil, com coeficiente de .47. Mais

uma vez, a correlação da moeda Real com a métrica calculada para as menções a todas as entidades obteve o melhor resultado (.52) quando comparado a cada entidade individualmente.

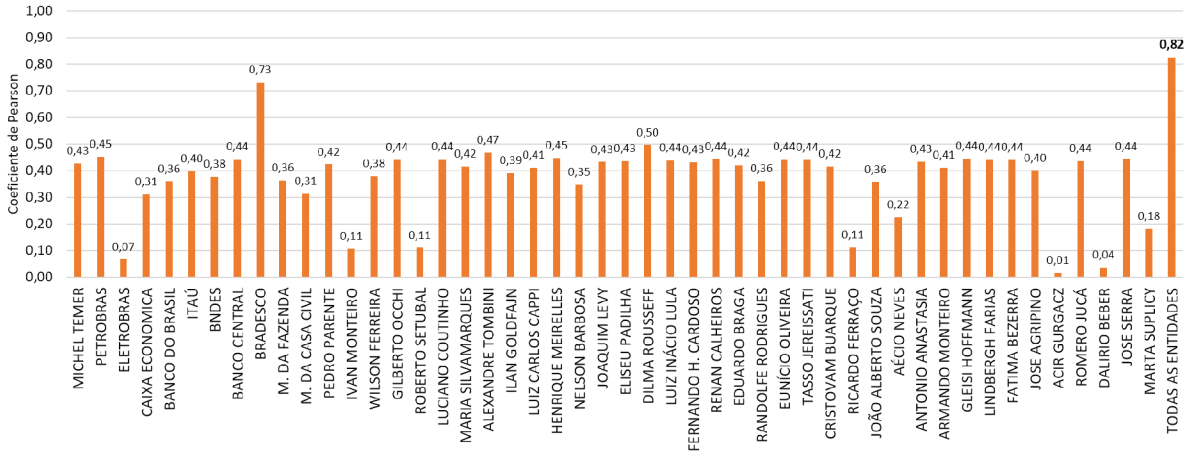
Figura 18 – Coeficiente de Pearson somente para entidades do tipo Organização



Fonte: elaborado pelo autor.

A Figura 18 exibe os coeficientes de correlação entre a moeda Real e todas as menções as entidades do tipo Organização. O gráfico mostra que menções ao Ministério da Fazenda e ao Ministério da Casa Civil, que se esperava ter uma grande influência na cotação da moeda Brasileira, tem baixa correlação com a cotação do Real (.36 e .31, respectivamente) do que organizações como o maior banco privado Brasileiro, o Bradesco, cuja a correlação (.73) pode ser considerada alta.

Figura 19 – Coeficiente de Pearson para todas as entidades



Fonte: elaborado pelo autor.

As Figuras 19 e 18 mostram que menções ao Bradesco apresentam a maior correlação com a cotação do Real dentre todas as entidades consideradas. Por outro lado,

algumas entidades (e.g., senador Dalirio Beber, senador Ricardo Ferraço, Roberto Setúbal que foi o presidente do banco Itaú) apresentaram coeficientes de correlação próximos a 0. As menções a essas entidades, relativamente sem importância, ainda podem contribuir para a correlação da métrica consolidada para todas as entidades consideradas, que foi calculada com os respectivos pesos da entidade (.82). Entretanto, embora esse coeficiente possa ser considerado alto, é inferior ao coeficiente de correlação com a métrica consolidada apenas para entidades do tipo Organização (.88).

Note que a métrica consolidada f_3 para um conjunto de entidades sempre apresenta melhor desempenho de correlação com a moeda Real do que a mesma métrica aplicada individualmente para cada entidade do respectivo conjunto. Porém considerar somente entidades do tipo Organização tem melhores resultados ao comparar com todas as entidades. Esses resultados sugerem que considerar apenas o número maior de entidades na métrica consolidada não garante uma melhor correlação. A seleção adequada de entidades a serem consideradas é crucial para maximizar a correlação.

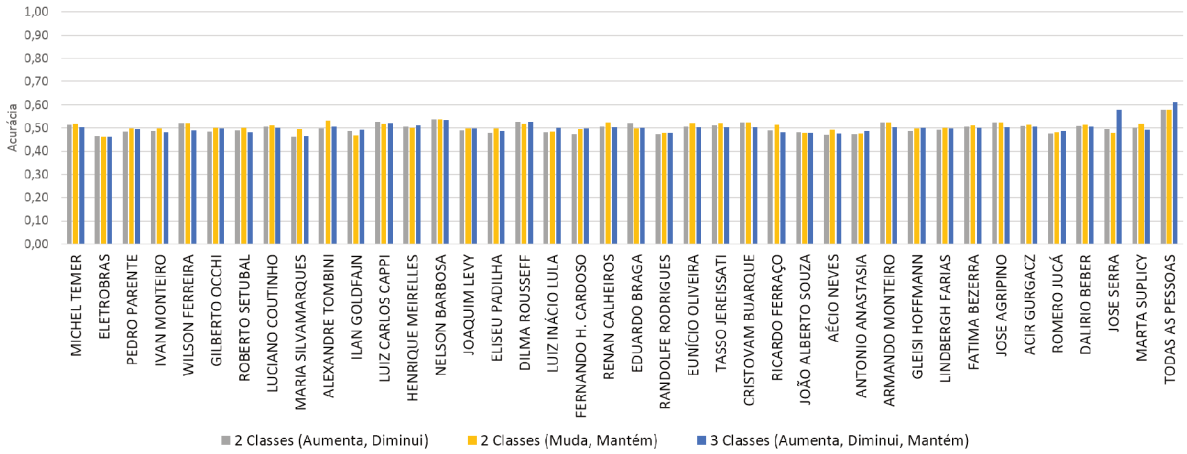
6.2.2 Análise Preditiva

As análises preditivas são de duas maneiras. A primeira maneira utiliza classificadores OneR (da Figura 20 até a Figura 25) para predizer as oscilações da moeda Real. A segunda maneira (da Figura 26 até a Figura 31), por sua vez, utiliza a técnica de regressão linear simples para predizer o valor da cotação da moeda. O uso de regressão linear permite verificar fenômenos de causa e efeito para predizer a cotação da moeda, ao invés de prever, apenas, as oscilações, como ocorre com os classificadores.

O eixo horizontal dos gráficos que compreendem os classificadores OneR representam as entidades analisadas nos respectivos gráficos e uma legenda indicando as cores usadas para representar os resultados para os classificadores distintos. As três últimas barras de cada um dos seis gráficos referem-se à predição da cotação do Real com a métrica consolidada f_3 para todas as entidades consideradas nos respectivos gráficos. O primeiro classificador, *2 Classes (Aumenta, Diminui)* (em cinza), prevê se a moeda irá aumentar ou diminuir. O segundo, *2 Classes (Muda, Mantém)* (amarelo), prediz se a cotação da moeda irá mudar ou não. Finalmente, o terceiro classificador, *3 Classes (Aumenta, Diminui, Mantém)* (azul), prediz se a moeda irá aumentar, diminuir ou permanecer estável.

A Figura 20 apresenta as acurácias das predições dos classificadores OneR para as entidades do tipo Pessoa. Nota-se que independente do classificador os valores de acurácias são em torno de .5, o que é um resultado considerado ruim. No classificador ternário (*3 Classes (Aumenta, Diminui, Mantém)*) somente a entidade José Serra conseguiu atingir uma acurácia de predição próxima a .6. Apesar dos resultados as entidades individualmente não serem bons, ao analisar todas as entidades

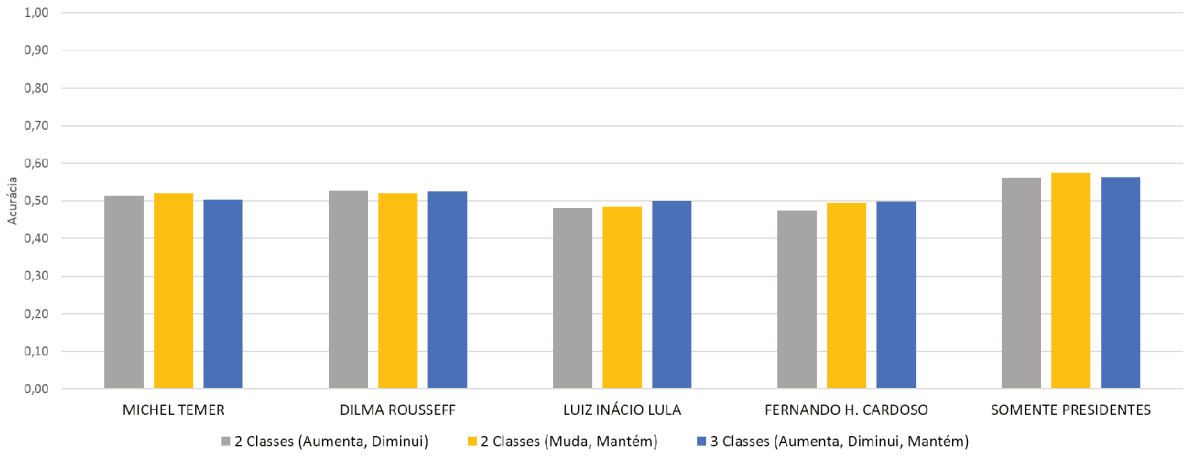
Figura 20 – Acurácias dos classificadores OneR para as entidades do tipo Pessoa



Fonte: elaborado pelo autor.

(últimas três barras do gráfico) é possível perceber a superioridade dessa métrica consolidada, pois os valores de acurácia em todos os classificadores são próximos a .6.

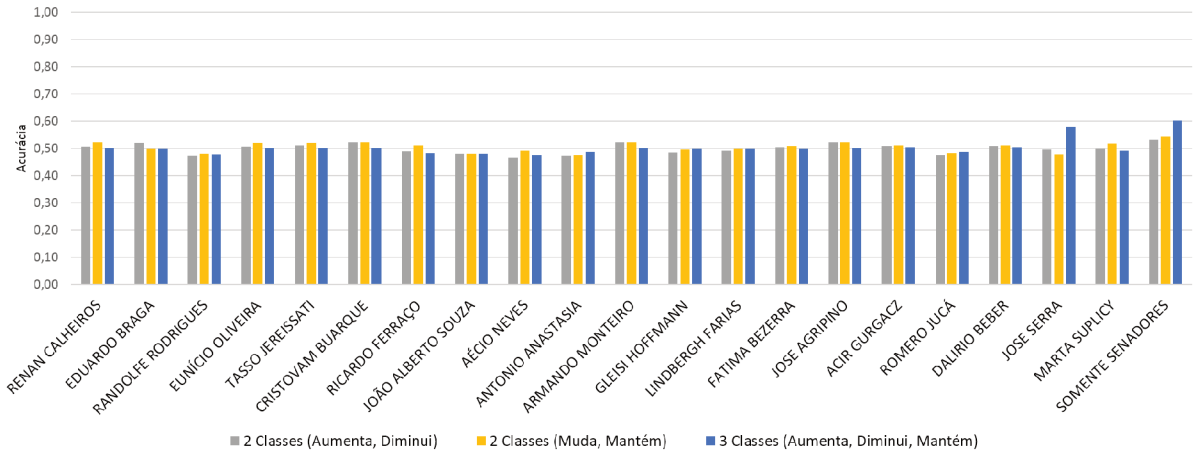
Figura 21 – Acurácias dos classificadores OneR para as entidades do tipo Pessoa com papel de Presidente



Fonte: elaborado pelo autor.

A Figura 21 mostra a acurácia da predição dos classificadores para entidades do tipo Pessoa com papel de Presidente. Todas essas acurácias estão abaixo de .6, o que é considerado baixo, com pequenos ganhos obtidos pela consolidação das métricas para todas as entidades. Quando se compara as acurácias obtidas pela métrica consolidada com a mesma métrica consolidada da figura anterior, se percebe uma diminuição de acurácia para o classificador ternário e uma sutil melhora no classificador binário *2 Classes (Muda, Mantém)*.

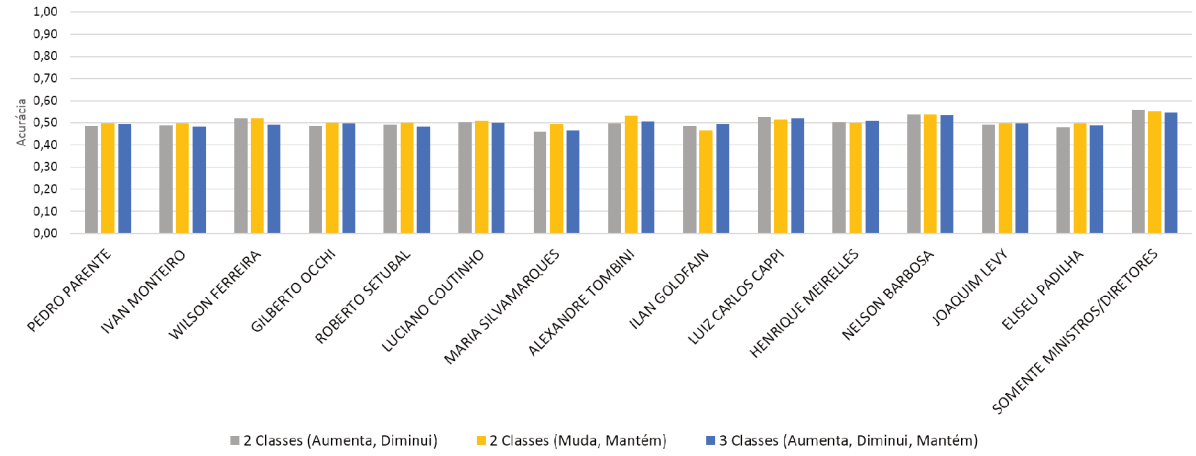
Figura 22 – Acurácias dos classificadores OneR para as entidades do tipo Pessoa com papel de Senador



Fonte: elaborado pelo autor.

A Figura 22 exibe as acurácias dos classificadores OneR para as entidades do tipo Pessoa com papel de Senador. O comportamento dos classificadores continua mantendo o mesmo comportamento, com valores de acurácias próximos a .5. A métrica consolidada com todos os senadores apresenta comportamento semelhante as acurácias da métrica consolidada para todas as entidades do tipo Pessoa, inclusive com o destaque para o classificador ternário com acurácia de .6.

Figura 23 – Acurácias dos classificadores OneR para as entidades do tipo Pessoa com papeis de Ministro ou Diretor

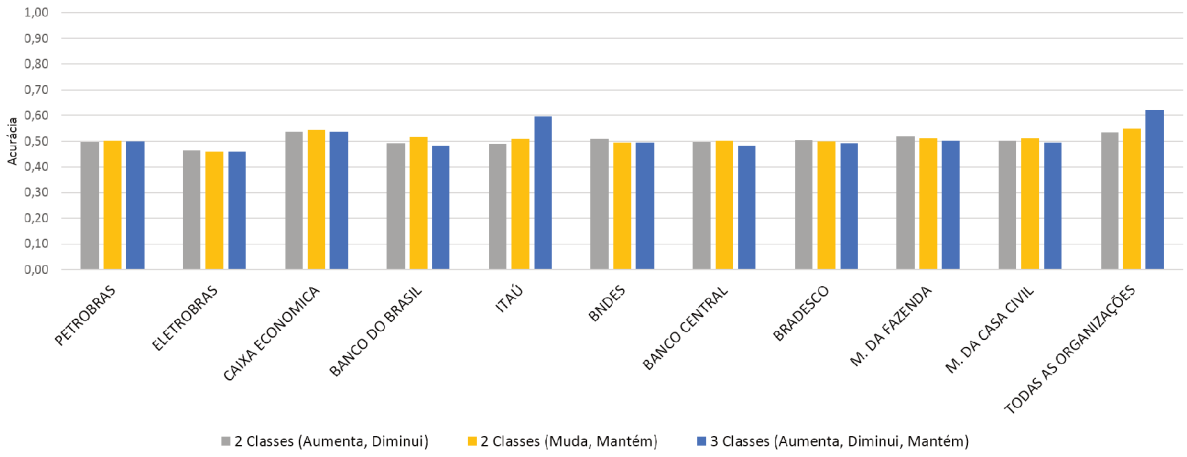


Fonte: elaborado pelo autor.

A Figura 23 apresenta as acurácias dos classificadores OneR para as entidades do tipo Pessoa com papeis de Ministro ou Diretor. O comportamento das acurácias mantém comportamento semelhante aos apresentados anteriormente. Todas as acurácias são menores que .6 com a maioria dos classificadores registrando acurácias de

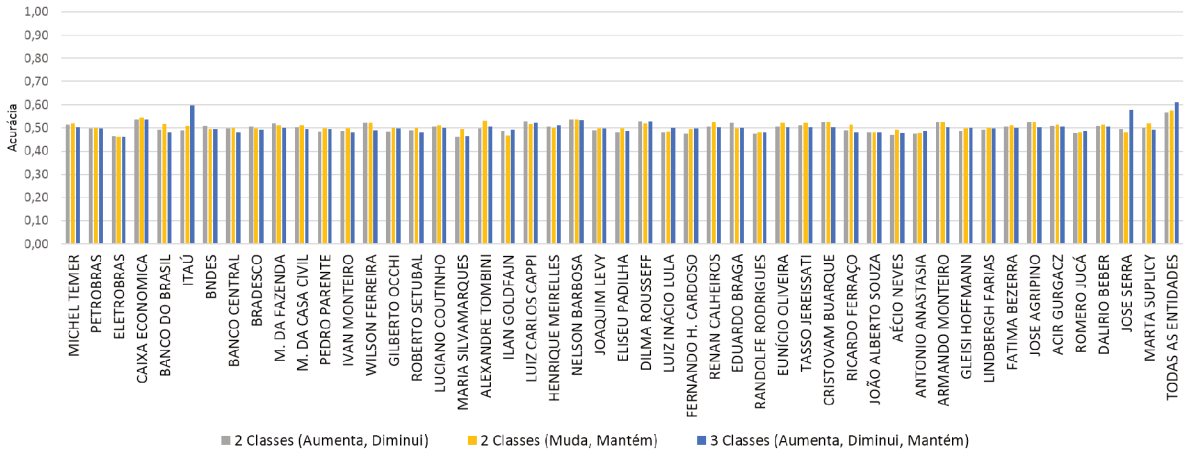
.5.

Figura 24 – Acurácias dos classificadores OneR para as entidades do tipo Organização



Fonte: elaborado pelo autor.

Figura 25 – Acurácias dos classificadores OneR para todas as entidades



Fonte: elaborado pelo autor.

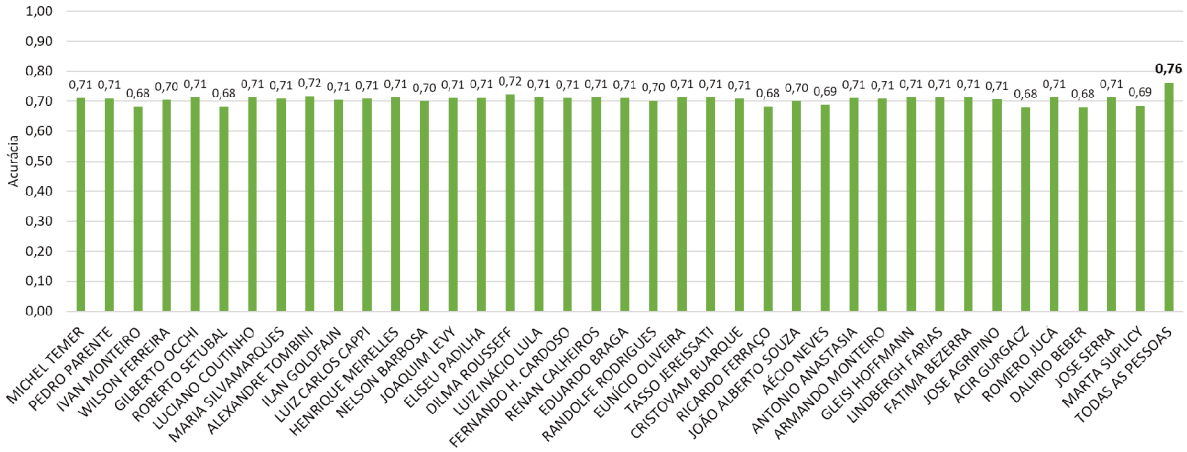
A Figura 24 mostra as acurácias das predições que foram obtidas considerando entidades do tipo Organização. Apenas a acurácia do classificador ternário considerando as menções ao Banco Itaú e todas as organizações consideradas chegam perto de .6. Finalmente, a Figura 25 mostra as acurácias obtidas considerando-se menções a todos os tipos de entidades consideradas nesses experimentos. Neste gráfico vale a pena notar a precisão próxima de .58 do classificador ternário ao considerar menções ao senador José Serra, que não se esperava que fosse mais influente do que outras entidades, da mesma forma que o Banco Itaú.

Observe que as acurácias das predições baseadas na métrica consolidada f_3 para um conjunto de entidades são sempre superiores a qualquer acurácia obtida

com a mesma métrica para cada entidade, individualmente, do conjunto. Note também que toda moeda tende a ter flutuação diariamente, porém os dias que não ocorrem oscilações são primordiais para diferenciar o classificador binário *2 Classes (Aumenta, Diminui)* do classificador ternário *3 Classes (Aumenta, Diminui, Mantém)*. Apesar dessas considerações, reforça-se que de em geral os classificadores obtiveram predições ruins, com valores máximos de .6.

As predições que utilizam a técnica de regressão linear simples (RLS) são apresentas a partir da Figura 26 até a Figura 31. O eixo horizontal dos gráficos presentes nessas figuras enumera as entidades consideradas para realizar as predições. A última barra de cada uma das seis figuras refere-se à predição da cotação da moeda Real com a métrica consolidada f_3 para todas as entidades consideradas nos respectivos gráficos.

Figura 26 – Acurácias das regressões lineares para as entidades do tipo Pessoa



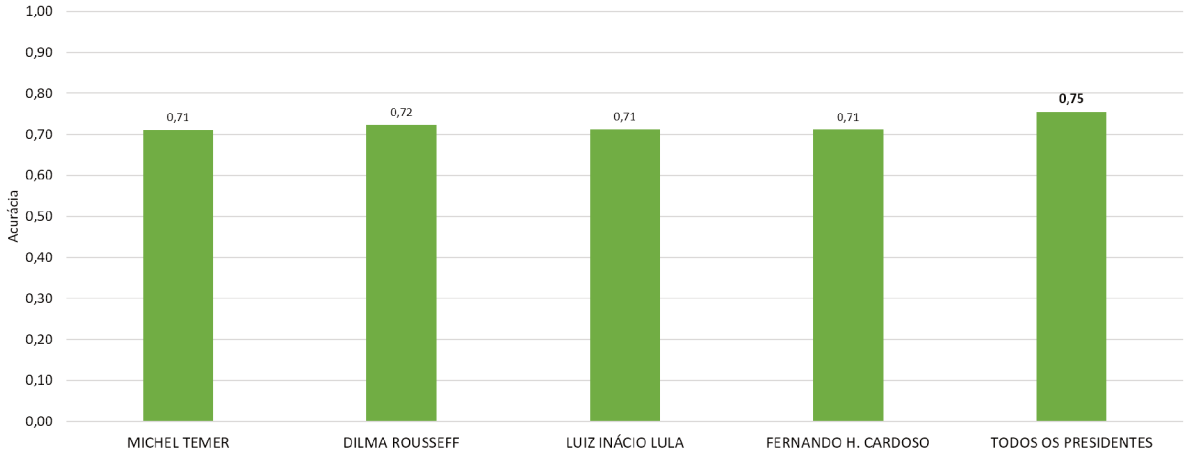
Fonte: elaborado pelo autor.

A Figura 26 apresenta as acurácias obtidas através da RLS da cotação da moeda Real e das menções as entidades do tipo Pessoa. As acurácias obtidas oscilam de .68 até .72 entre as entidades individualmente. Esses resultados podem ser considerados satisfatórios, quando comparado aos resultados obtidos pelos classificadores OneR. Já a acurácia obtida ao se calcular a métrica consolidada para todas as entidades do tipo Pessoa foi de .76.

A Figura 27 mostra os resultados obtidos considerando as menções as entidades do tipo Pessoa desempenhando o papel de Presidente. Considerando as entidades individualmente, as acurácias estavam entre .71 e .72, enquanto a métrica consolidada para todas essas pessoas apresenta um leve ganho (.75) porém inferior ao obtido pela métrica consolidada para todas as entidades do tipo Pessoa somente.

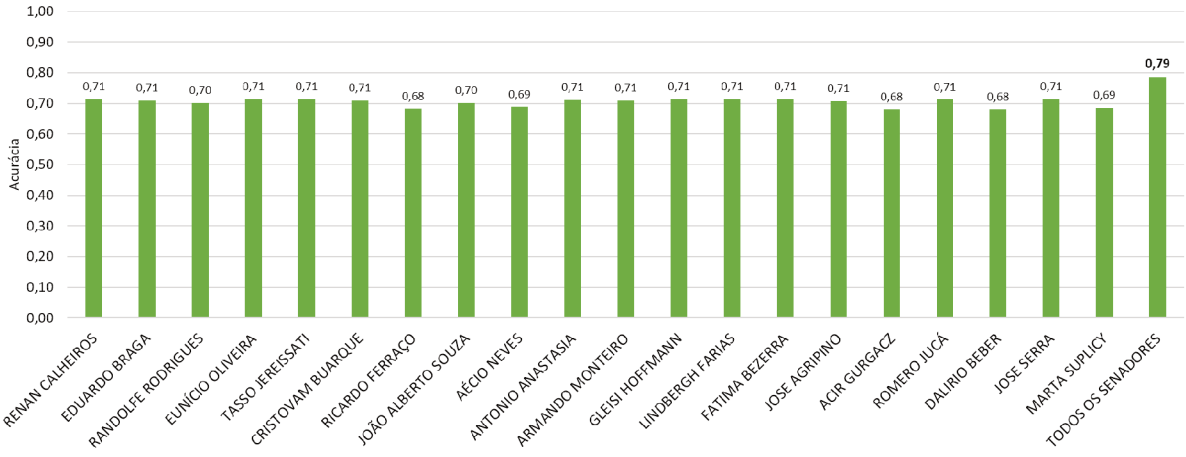
A Figura 28 exibe as acurácias obtidas através da RLS da cotação da moeda Real e das menções as entidades do tipo Pessoa com papel de Senador. As acurácias,

Figura 27 – Acurácias das regressões lineares para as entidades do tipo Pessoa com papel de Presidente



Fonte: elaborado pelo autor.

Figura 28 – Acurácias das regressões lineares para as entidades do tipo Pessoa com papel de Senador



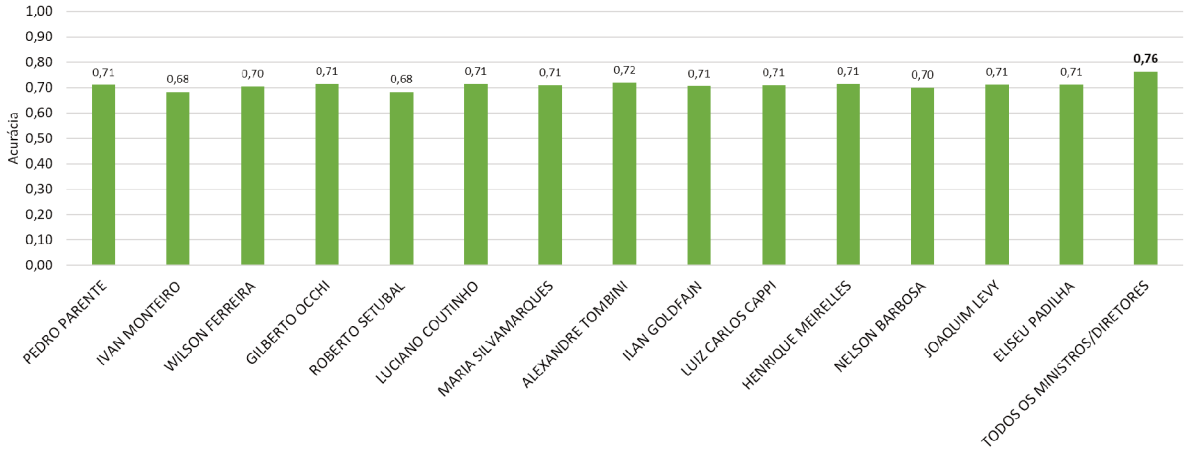
Fonte: elaborado pelo autor.

mais uma vez, oscilaram entre .68 e .71 para todas as entidades individualmente, porém alcançou .79 para a métrica consolidada para todas as entidades em questão.

A Figura 29 exibe as acurácias obtidas através da RLS da cotação da moeda Real e das menções as entidades do tipo Pessoa com papeis de Ministro ou Diretor. O comportamento das acurácias segue semelhantes aos apresentados até então. A métrica consolidada para todas as entidades registrou acurácia de .76.

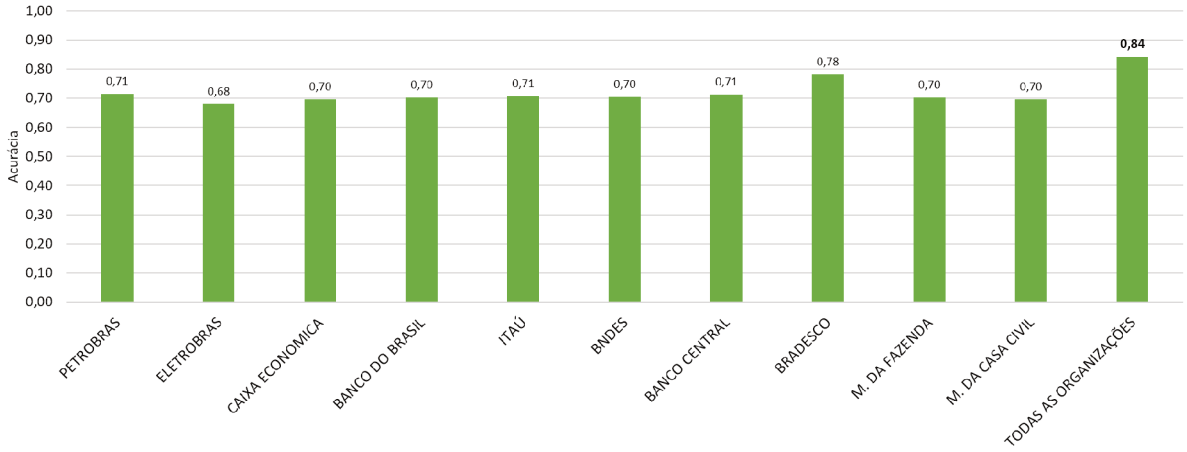
A Figura 30, por sua vez, apresenta os resultados obtidos considerando-se menções a entidades do tipo Organização. As entidades aqui representadas apresentam comportamento similares, com acurácias próxima a .70, exceto a entidade Bradesco, com acurácia de .78. Já a métrica consolidada para todas as entidades registrou acu-

Figura 29 – Acurácias das regressões lineares para as entidades do tipo Pessoa com papeis de Ministro ou Diretor



Fonte: elaborado pelo autor.

Figura 30 – Acurácias das regressões lineares para as entidades do tipo Organização



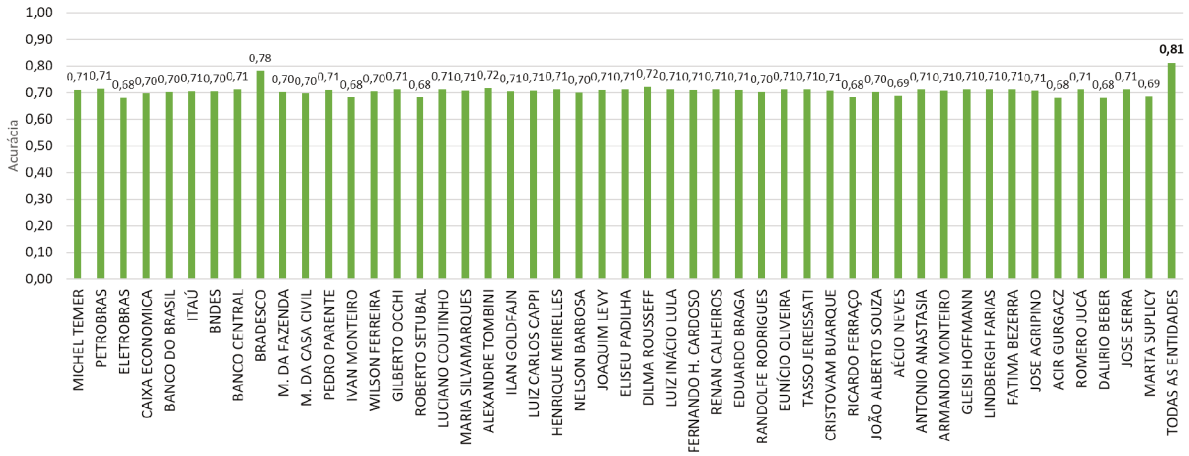
Fonte: elaborado pelo autor.

rácia de .84, um resultado bom dentre todos até aqui obtidos.

Finalmente, a Figura 31 mostra as acurácias obtidas individualmente para cada entidade considerada nos experimentos. As entidades aqui presentes foram apresentadas nos outros gráficos, exceto pela métrica consolidada. A acurácia da métrica consolidada alcança .81, um bom resultado, porém inferior ao obtido pela métrica consolidada ao se considerar somente entidades do tipo Organização.

Nota-se nesses gráficos que menções às organizações financeiras privada, particularmente Itaú e Bradesco, levam aos melhores resultados individuais neste estudo de caso, mostrando sua influência sobre as oscilações da moeda brasileira. Este estudo de caso também sugere que considerar um número maior de entidades não leva necessariamente a melhores resultados, como se pode observar comparando

Figura 31 – Acurácias das regressões lineares para todas as entidades



Fonte: elaborado pelo autor.

os resultados nas Figuras 30 e 31.

As dificuldades para obter predições mais precisas com os classificadores OneR provavelmente se devem ao fato de que a cotação da moeda é um valor quantitativo contínuo. Desta maneira, foi necessário analisar a variação da moeda em termos de períodos de tempo e convertê-la em valores qualitativos. Além disso, a precisão das cotações da moeda torna a configuração apropriada de classificadores bastante sutil, devido a pequenas oscilações (e.g. exemplo, de 2.987 para 2.986), por exemplo. A regressão linear simples, por outro lado, que utiliza variáveis quantitativas contínuas, não requer tais conversões. Isso leva a melhores acurácias, entre .68 e .78 ao considerar entidades individualmente e até .84 usando a métrica consolidada para todas as entidades do componente conectado do SN usado em nossos experimentos.

7 CONCLUSÕES

Esse trabalho propôs o método *5-Ions* que utiliza ferramentas de PLN para extrair automaticamente características de texto (sumários, menções a entidades nomeadas e sentimentos associados) para estimar o desempenho de uma entidade alvo com base em métricas calculadas com as características extraídas de textos. A principal inovação do método proposto é a utilização de redes semânticas (RS) para calcular métricas de desempenho consolidadas para entidades semanticamente relacionadas. Isso permite explorar todas as menções a entidades semanticamente relacionadas a uma entidade alvo (i.e. entidades que estão em um mesmo componente conexo da RS que contém a entidade alvo) para estimar o seu desempenho real. O uso de métricas de desempenho consolidadas para entidades semanticamente relacionadas pode compensar um baixo índice de menções a alguma entidade alvo. Foi desenvolvido um algoritmo genético para definir os pesos das entidades em métricas de desempenho consolidadas, de forma a maximizar o desempenho das correlações de tais métricas com medidas reais de desempenho e a acurácia das predições obtidas usando tais métricas consolidadas. O trabalho também analisou o impacto da sumarização automática de texto e de correferências no desempenho das correlações e predições.

Os experimentos para avaliar o desempenho do método proposto analisou dois estudos de caso. O primeiro estudo de caso estimou a variação da intenção de voto dos partidos políticos Alemão, analisando também as menções aos afiliados políticos dos respectivos partidos. O uso de métricas consolidadas para as entidades semanticamente relacionadas obteve as melhores correlações e as maiores acurácias das predições em comparação aos estimativas que considera apenas menções as entidades alvo, em todos os cenários experimentados. Os coeficientes de correlação entre as medidas reais de desempenho e as métricas consolidadas para as entidades relacionadas sempre foram melhores ou equivalentes àqueles com métricas levando em consideração apenas as menções as entidades alvo. Para as predições, foram obtidos ganhos que variaram de 15% a 89%, usando métricas de desempenho consolidadas em vez de métricas para apenas as entidades alvo. A utilização de sumarização neste estudo de caso trouxe ganhos em tempo de processamento de até 31% no tempo total de extração das características textuais. Porém as correlações de medidas de desempenho com métricas calculadas a partir de textos sumarizados apresentaram oscilações de ganhos e perdas, em comparação com as métricas calculadas a partir dos textos sem sumarização. Já as predições apresentaram resultados satisfatórios mesmo com métricas calculadas a partir de textos sumarizados, com mais ganhos que perdas, quando comparadas as acurácias obtidas a partir dos textos não sumarizados. Esses resultados sugerem que melhores técnicas de sumarização podem apresentar ganhos ao processo.

O segundo estudo de caso analisou a valorização da moeda corrente no Brasil, o Real, através da análise de menções a políticos e instituições públicas e privadas de relevância para a economia nacional. Os experimentos utilizaram técnicas de aprendizado de máquina, similares às do primeiro estudo de caso, e regressão linear simples. A entidade alvo, Real, não teve suas menções monitoradas, uma vez que os textos analisados apresentaram baixo número de menções a ela, ficando a cargo das outras entidades serem correlacionadas e predizerem o seu desempenho. A correlação da instituição financeira Bradesco alcançou o coeficiente de .73, porém ao se considerar as métricas consolidas os resultados chegaram a .88. As predições de variações (i.e, aumenta, diminui, mantém), por outro lado, não obtiveram resultados tão satisfatórios. Porém, ao se aplicar técnicas de regressão linear (há correlação entre as medidas) os resultados de predição alcançaram acurácia de .84, resultado que é considerado muito bom. O segundo estudo de caso não analisou o impacto das sumarizações de textos.

As principais conclusões obtidas desse trabalho são: (i) os ganhos de tempo de processamento ao aplicarmos sumarização automática de texto são significativos, porém os benefícios quanto a análise de correlação e predição são inconclusivos, necessitando de um maior aprofundamento; (ii) a utilização de classificadores para realização de predição é viável em alguns casos, porém outros casos exigem técnicas mais específicas de predição, por exemplo, a regressão linear simples aplicada para a flutuação da moeda; (iii) por fim, a utilização de redes semânticas gera ganhos de correlação e predição ao se analisar uma entidade alvo, considerando menções a essa entidade ou não, porém o maior número de entidades na rede semântica não implica, necessariamente, em melhores resultados, sendo necessário selecionar corretamente as entidades que compõe a rede.

Esta pesquisa resultou em três artigos científicos. O primeiro artigo descreve o método *5-Ions* e aborda o estudo de caso da intenção de voto, porém não realiza a comparação dos textos originais com os textos sumarizados. Esse primeiro artigo foi aceito na *21st International Conference on Information Integration and Web-based Applications & Services (iiWAS 2019)* e será publicado em dezembro de 2019. O segundo artigo apresenta o estudo de caso da flutuação da moeda, mostrando ser possível correlacionar e predizer o desempenho de uma entidade alvo monitorando apenas entidades a ela relacionada. Esse artigo inclui a aplicação de técnicas de regressão para predição de desempenho. Ele foi aceito no Simpósio Brasileiro de Sistemas de Informação (SBSI 2020) e será publicado em maio de 2020. Por fim, o terceiro artigo está sendo preparado e conta a exploração da sumarização de texto, além da aplicação de diferentes técnicas de regressão e aprendizado de máquina para predizer o desempenho das entidades alvo. O alvo de submissão desse artigo ainda será definido.

A adoção de ferramentas prontas de PLN para extrair características de texto,

tal como entidades nomeadas e sentimentos, e a exploração de redes semânticas para produzir indicadores e métricas de desempenho consolidadas, usando técnicas de busca de soluções como algoritmos genéticos, além de regressão para medidas de desempenho reais contínuas, abre novas perspectivas para a análise de desempenho e a predição de desempenho de entidades com base em suas menções e sentimentos associados, em uma variedade de textos produzidos na Web. Temas para trabalhos futuros incluem: (i) desenvolvimento e teste de novas métricas de desempenho consolidadas para entidades semanticamente relacionadas, aplicando métodos de inteligência artificial para identificar as métricas com melhores desempenhos; (ii) examinar o comportamento de outros classificadores e métodos para realizar a predição de desempenho, incluindo combinação de classificadores; e (iii) uma melhor investigação do impacto da utilização de ferramentas do estado da arte em NER/NED, análise de sentimentos e sumarização de textos no *5-Ions*, visando ganhos de precisão e cobertura dessas tarefas, enquanto se mantém o tempo de execução sob controle para grandes volumes de dados, com consequentes melhorias nas correlações das métricas produzidas a partir dos textos com medidas de desempenho real e na própria predição de desempenho.

REFERÊNCIAS

AHMED, Saifuddin; SKORIC, Marko M. My Name Is Khan: The Use of Twitter in the Campaign for 2013 Pakistan General Election. *In*: PROCEEDINGS of the 2014 47th Hawaii International Conference on System Sciences. Washington, DC, USA: IEEE Computer Society, 2014. (HICSS '14), p. 2242–2251. DOI: 10.1109/HICSS.2014.282. Disponível em: http://dx.doi.org/10.1109/HICSS.2014.282.

ARVATE, Paulo Roberto. PolÃtica cambial no Brasil: um esquema analÃtico. pt. **Relatório de pesquisa FGV/EAESP/NPP**, FGV, v. 5, 2004. Disponível em: http://bibliotecadigital.fgv.br/dspace/handle/10438/3186?show=full.

BENESTY, J. *et al.* Pearson Correlation Coefficient. *In*: NOISE Reduction in Speech Processing. [*S.l.*]: Springer, 2009. v. 2. ISBN 978-3-642-00295-3.

BONTCHEVA, Kalina; ROUT, Dominic. Making sense of social media streams through semantics: A survey. **Semantic Web**, 2012. ISSN 1570-0844. DOI: 10.3233/SW-130110. Disponível em: http://dx.doi.org/10.3233/SW-130110.

CIRQUEIRA, Douglas *et al.* Performance Evaluation of Sentiment Analysis Methods for Brazilian Portuguese. *In*: ABRAMOWICZ, Witold; ALT, Rainer; FRANCZYK, Bogdan (Ed.). **Business Information Systems Workshops**. Cham: Springer International Publishing, 2017. p. 245–251.

DAS, Dipanjan; MARTINS, André F. T. **A Survey on Automatic Text Summarization**. [*S.l.*: *s.n.*], 2007.

DEERWESTER, Scott *et al.* Indexing by latent semantic analysis. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE**, v. 41, n. 6, p. 391–407, 1990.

FILETO, Renato *et al.* A Semantic Model for Movement Data Warehouses. *In*: PROCEEDINGS of the 17th International Workshop on Data Warehousing and OLAP, {DOLAP} 2014, Shanghai, China, November 3-7, 2014. [*S.l.*: *s.n.*], 2014. p. 47–56. DOI: 10.1145/2666158.2666180. Disponível em: http://doi.acm.org/10.1145/2666158.2666180.

FINK, Clay *et al.* Twitter, Public Opinion, and the 2011 Nigerian Presidential Election. **Proceedings of the 2014 47th Hawaii International Conference on System Sciences**, IEEE Computer Society, set. 2013. DOI: 10.1109/SocialCom.2013.50. Disponível em: https://doi.org/10.1109/SocialCom.2013.50.

FRANCIA, Matteo; GOLFARELLI, Matteo; RIZZI, Stefano. A methodology for social BI. *In*: ACM. PROCEEDINGS of the 18th International Database Engineering & Applications Symposium. [*S.l.*: *s.n.*], 2014. p. 207–216.

GRZEGORZEWSKI, Przemyslaw; ZIEMBINSKA, Paulina. Spearman's Rank Correlation Coefficient for Vague Preferences. *In*: FLEXIBLE Query Answering Systems. [*S.l.*]: Springer, 2011. ISBN 978-3-642-24763-7.

KUMAR, Shamanth; MORSTATTER, Fred; LIU, Huan. **Twitter Data Analytics**. [*S.l.*]: Springer Publishing Company, Incorporated, 2013. ISBN 1461493714.

LINDEN, R. **Algoritmos Genéticos (2a edição)**. [*S.l.*]: BRASPORT, 2008. ISBN 9788574523736. Disponível em: https://books.google.com.br/books?id=it0kv6UsEMEC.

LIU, Bing. **Sentiment Analysis and Opinion Mining**. [*S.l.*]: Morgan & Claypool Publishers, 2012. ISBN 1608458849.

LUHN, H. P. The Automatic Creation of Literature Abstracts. **IBM Journal of Research and Development**, v. 2, n. 2, p. 159–165, abr. 1958. DOI: 10.1147/rd.22.0159. Disponível em: http://ieeexplore.ieee.org/document/5392672/.

MAHAJAN, Aditi; GANPATI, Anita. Performance Evaluation of Rule Based Classification Algorithms. **International Journal of Advanced Research in Computer Engineering and Technology**, v. 3, n. 10, 2014.

MENDES, Pablo N *et al.* DBpedia-Spotlight: shedding light on the web of documents. *In*: ACM. PROCEEDINGS of the 7th international conference on semantic systems. [*S.l.*: *s.n.*], 2011. p. 1–8.

MITCHELL, Melanie. **An Introduction to Genetic Algorithms**. Cambridge, MA, USA: MIT Press, 1998. ISBN 0262631857.

MORO, Andrea; RAGANATO, Alessandro; NAVIGLI, Roberto. Entity linking meets word sense disambiguation: a unified approach. **Transactions of the Association for Computational Linguistics**, v. 2, p. 231–244, 2014.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Linguisticae Investigationes**, v. 30, n. 1, p. 3–26, jan. 2007. Publisher: John Benjamins Publishing Company. Disponível em: http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002.

NEBOT, Victoria; BERLANGA, Rafael. Building data warehouses with semantic web data. **Decision Support Systems**, Elsevier, v. 52, n. 4, p. 853–868, 2012.

PATTEKARI, S.A.; PARVEEN, A. Prediction system for heart disease using naive bayes. v. 3, p. 290–294, jan. 2012.

PEREIRA, Vilmar César Júnior *et al.* A Semantic BI Process for Detecting and Analyzing Mentions of Interest for a Domain in Tweets. *In*: PROCEEDINGS of the 24th

Brazilian Symposium on Multimedia and the Web. Salvador, BA, Brazil: ACM, 2018. (WebMedia '18), p. 197–204. DOI: 10.1145/3243082.3243100. Disponível em: http://doi.acm.org/10.1145/3243082.3243100.

PRADHAN, Sameer *et al.* CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. *In*: PROCEEDINGS of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Portland, Oregon: Association for Computational Linguistics, 2011. (CONLL Shared Task '11), p. 1–27. Disponível em: http://dl.acm.org/citation.cfm?id=2132936.2132937.

RAMTEKE, Jyoti *et al.* Election result prediction using Twitter sentiment analysis. **Inventive Computation Technologies (ICICT), International Conference on**, IEEE Computer Society, ago. 2016. DOI: 10.1109/INVENTIVE.2016.7823280. Disponível em: https://doi.org/10.1109/INVENTIVE.2016.7823280.

RAO, Delip; MCNAMEE, Paul; DREDZE, Mark. Entity Linking: Finding Extracted Entities in a Knowledge Base. *In*: POIBEAU, Thierry *et al.* (Ed.). **Theory and Applications of Natural Language Processing**. [*S.l.*]: Springer, 2013. cap. 4.

REVELLE, William. **An introduction to psychometric theory with applications in R**. [*S.l.*: *s.n.*], 2013.

RIBEIRO, Haroldo V *et al.* The dynamical structure of political corruption networks. **Journal of Complex Networks**, cny002, 2018. DOI: 10.1093/comnet/cny002. eprint: /oup/backfile/content_public/journal/comnet/pap/10.1093_comnet_cny002/1/cny002.pdf. Disponível em: http://dx.doi.org/10.1093/comnet/cny002.

ROSSI, PEDRO. PolÃtica cambial no Brasil: um esquema analÃtico. pt. **Brazilian Journal of Political Economy**, scielo, v. 35, p. 708–727, dez. 2015. ISSN 0101-3157. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-31572015000400708&nrm=iso.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. 2. ed. [*S.l.*]: Pearson Education, 2003. ISBN 0137903952.

SACENTI, Juarez AP *et al.* Automatically tailoring semantics-enabled dimensions for movement data warehouses. *In*: SPRINGER. INTERNATIONAL Conference on Big Data Analytics and Knowledge Discovery. [*S.l.*: *s.n.*], 2015. p. 205–216.

SARANYAMOL, C S; SINDHU, L. **A Survey on Automatic Text Summarization**. v. 5. [*S.l.*: *s.n.*], 2014.

SHEN, W.; WANG, J.; HAN, J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. **IEEE Transactions on Knowledge and Data Engineering**, v. 27, n. 2, p. 443–460, fev. 2015. ISSN 1041-4347. DOI: 10.1109/TKDE.2014.2327028.

SHEN, Wei; WANG, Jianyong; HAN, Jiawei. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE, p. 443–460, 2014. DOI: 10.1109/TKDE.2014.2327028. Disponível em: https://doi.org/10.1109/TKDE.2014.2327028.

SOWA, John F. (Ed.). **Principles of Semantic Networks: Explorations in the Representation of Knowledge**. [*S.l.*]: Morgan Kaufmann, 1991. (The Morgan Kaufmann Series in Representation and Reasoning).

SPECK, René; NGOMO, Axel-Cyrille Ngonga. Named entity recognition using fox. *In*: CEUR-WS. ORG. PROCEEDINGS of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272. [*S.l.*: *s.n.*], 2014. p. 85–88.

TUMITAN, D.; BECKER, K. Sentiment-based Features for Predicting Election Polls: A Case Study on the Brazilian Scenario. **Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM Intl. Joint Conferences on**, IEEE Computer Society, Warsaw, Poland, ago. 2014. DOI: 10.1109/WI-IAT.2014.89. Disponível em: http://ieeexplore.ieee.org/document/6927616/.

VANDERWENDE, Lucy *et al.* Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. **Inf. Process. Manage.**, v. 43, n. 6, p. 1606–1618, 2007.

VIEIRA, Renata; LIMA, Vera Lúcia Strube. Lingüística Computacional: princípios e aplicações. **IX Escola de Informática da SBC-Sul**, v. 30, p. 1–42, 2001.

VILLANUEVA, Daniel *et al.* SMORE: Towards a semantic modeling for knowledge representation on social media. **Science of Computer Programming**, Elsevier, v. 121, p. 16–33, 2016.

XU, Weichao *et al.* A Comparative Analysis of Spearman's Rho and Kendall's Tau in Normal and Contaminated Normal Models. **Signal Process.**, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 93, n. 1, p. 261–276, jan. 2013. ISSN 0165-1684. DOI: 10.1016/j.sigpro.2012.08.005. Disponível em: http://dx.doi.org/10.1016/j.sigpro.2012.08.005.

YAN, Xin; SU, Xiao Gang. **Linear Regression Analysis: Theory and Computing**. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2009. ISBN 9789812834102.