Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Análise de Fluxo de Aspectos Linguísticos em Textos: uma Abordagem Inspirada na Análise de Áudio

Larissa Lucena Vasconcelos

Proposta de Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Processamento de Linguagem Natural

Claudio Elízio Calazans Campelo
(Orientador)

Campina Grande, Paraíba, Brasil

Resumo

Um desafio enfrentado pela área de Processamento de Linguagem Natural (PLN) é representar e analisar textos de forma a reproduzir como humanos percebem aspectos linguísticos (subjetividade ou argumentação, por exemplo) neles presentes. Entender como esses aspectos são explorados pode levar a um melhor conhecimento sobre como diferentes tipos de texto são geralmente escritos. A correta interpretação de um texto, captando as nuances trazidas pela linguagem, é dada através da leitura do mesmo por completo. Logo, ao almejar entender como um aspecto linguístico é explorado ao longo de um texto, é necessário que o método utilizado para representação e análise do mesmo reflita o comportamento do aspecto por toda extensão do texto. Para isso, é preciso que seja conservada a forma como o aspecto foi abordado desde o início até o fim do texto. Dessa forma, este trabalho propõe um novo método de representação e análise que objetiva conservar a forma como um aspecto é explorado ao longo de todo o texto. O método proposto representa os textos através de fluxos de aspectos linguísticos, visando capturar e conservar o comportamento do aspecto ao longo dos mesmos. Ademais, inspirada na área de análise de áudio, a análise adapta conceitos como fragmentação dos fluxos em *frames* e também a extração de *features* inerentes do estudo dos áudios. Para validar a abordagem proposta, foram realizadas várias tarefas de classificação envolvendo diferentes línguas e aspectos linguísticos. A eficácia revelada pelos resultados obtidos indica a viabilidade do uso do método para representar e analisar textos em busca de refletir o comportamento dos aspectos, permitindo a aquisição de valioso conhecimento acerca dos textos.

Abstract

A challenge in Natural Language Processing is representing and analysing texts aiming to reproduce the way humans perceive linguistic aspects (e.g., subjectivity or argumentation) present in them. Understanding how these aspects are exploited can lead to better knowledge about how humans commonly write different kinds of text. The correct interpretation of a text, capturing the nuances brought by the language, is given by reading it thoroughly. Therefore, to achieve a better understanding of how a linguistic aspect is explored throughout a text, its representation and analysis must reflect the aspect's behavior all text long. It is then necessary to preserve how the text approaches an aspect from the beginning until the end. Thus, this work proposes a new method of representation and analysis that aims to maintain an aspect's behaviour throughout the text. The proposed method represents the texts through linguistic aspects flows, aiming to capture and preserve the aspect's behavior throughout them. Besides, inspired by the area of audio analysis, the analysis adapts concepts like fragmentation of flows into frames and also the extraction of features inherent in the audios' study. Several classification tasks, involving different languages and linguistic aspects, were carried out to initially validate the proposed approach. The effectiveness revealed by the results obtained indicates the feasibility of using the method to represent and analyze texts aiming to reflect the behavior of the aspects, allowing the acquisition of valuable knowledge about the texts.

Agradecimentos

Agradeço a Deus, primeiramente, por ter permitido o meu curso até esta etapa da vida, me concedendo saúde, inspiração e determinação para realização desta pesquisa.

Agradeço especialmente à minha mãe, Elza, minha maior inspiração e parceira de todas as "aventuras", pelo apoio dado em todos aspectos da minha vida e por sempre ter enfatizado o valor da dedicação em tudo que nos propomos a fazer. Agradeço também à Eveline, minha irmã e Isadora, sobrinha, que complementam essa família que é meu porto seguro.

Um agradecimento especial ao meu orientador, Cláudio Campelo, por toda paciência, confiança, parceria e dedicação dispensada a mim e a esta pesquisa. Obrigada por conduzir esta orientação exatamente da forma que eu precisava. Os ensinamentos que absorvi serão levados para a vida.

Meus agradecimentos também aos colegas do laborátorio Lacina, em especial a Caio Jerônimo, por todas as discussões, opiniões e contribuições que engrandeceram esta pesquisa.

Agradeço ao Instituto Federal da Paraíba - IFPB, instituição da qual sou professora efetiva, que propiciou minha dedicação exclusiva a este trabalho até então, concedendo afastamento para qualificação.

Agradeço também a todos que fazem a Coordenação do Programa de Pós-Graduação em Computação, por todo suporte dispensado quando necessário.

Enfim, agradeço a todos os familiares e amigos que torceram e me dispensaram qualquer forma de apoio durante esta jornada.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Um desafio enfrentado pela área de Processamento de Linguagem Natural (PLN) é reproduzir a forma como humanos percebem uma certa característica linguística (referida neste trabalho por aspecto linguístico, ou apenas aspecto) expressada em um texto e, com isso, atingir um melhor entendimento sobre como diferentes tipos de texto são escritos. Por exemplo, analisar como notícias falsas fazem uso da linguagem subjetiva (aspecto) pode levar a um conhecimento significativo sobre elas [29].

Sabe-se que, para um texto ser corretamente interpretado, é essencial que seja lido por completo. Tomando como exemplo a avaliação de um produto, o autor exprime sua opinião ao longo do texto explorando o aspecto sentimento, como é possível verificar em trabalhos na área de *Opinion Mining*, como, por exemplo, o trabalho de *Pang* e *Lee* [43]. Para construir a opinião, o autor pode usar diversas estratégias como elogiar algumas características (sentimento positivo) e criticar outras (sentimento negativo), ou, até mesmo, apresentar alguns elogios e apenas uma crítica mais contundente que o levaria a avaliar mal o produto ao final [36]. Em ambos os casos, o leitor vai captar de forma correta a opinião sobre o produto ao ponderar sobre cada um dos sentimentos apresentados no decorrer do texto.

Logo, para observar as particularidades de como um aspecto é apresentado em um texto, é necessária não apenas uma análise global do texto ou uma análise de cada palavra ou sentença individualmente, mas também uma análise de como o aspecto é explorado ao longo de todo o texto.

Entretanto, em PLN, ainda se observa uma predominância de estudos que se utilizam de técnicas que modelam representações locais de um aspecto para extrair *features* sintetizado-

ras que representam o texto como um todo, frequentemente por uma média ou mediana de representações locais, como, por exemplo, inferir o sentimento prevalescente em uma avaliação de produto a partir da média dos sentimentos identificados nas sentenças do texto. Conforme discutido por Aker et al. [3], esse tipo de representação pode levar à perda de informações importantes, especialmente para textos grandes, uma vez que podem ignorar singularidades do aspecto analisado presente em qualquer parte do texto que poderiam ser decisivas na identificação ou caracterização daquele tipo de texto.

Apresentada em trabalhos como o de Mao e Lebanon [36] e o de Wachsmuth e Stein [58], uma abordagem para evitar representar um texto de forma sintetizada e global é modelar o mesmo como um fluxo, uma sequência de informações coletadas a partir de palavras, sentenças ou parágrafos de um texto. Explorar as características dos fluxos, como a sequência das informações, ao realizar a análise dos textos, pode permitir uma análise mais eficaz sobre os textos estudados [17; 51; 34]. Por exemplo, como discute o trabalho de Jerônimo et al. [29], notícias falsas se utilizam com mais intensidade do aspecto subjetividade que notícias legítimas. Entretanto, já que o objetivo das notícias falsas é se passar por legítimas, pode haver partes do texto que explore a subjetividade de forma mais branda e mais parecida com as notícias legítimas. Se, ao analisar textos, for possível capturar essas nuances da variação do comportamento de um aspecto (e.g., por meio de fluxos), pode ser mais fácil caracterizar e diferenciar diversos tipos de texto, levando a um melhor conhecimento acerca deles.

Diante do exposto, o problema abordado nesta pesquisa de Doutorado é como representar e analisar textos de forma a extrair informações mais relevantes para entender como aspectos linguísticos são explorados ao longo deles, almejando se aproximar mais da forma como humanos percebem esses aspectos presentes no texto.

Conseguir analisar grandes volumes de texto identificando a forma como autores abordam certos aspectos pode impulsionar os resultados de tarefas como, por exemplo, classificação de notícias falsas, detecção de sarcasmo, classificação de sentimentos (opinião) sobre serviços ou produtos, avaliação de redações em concursos e análise de discursos políticos.

1.1 Objetivos e Questões de Pesquisa

O principal objetivo deste trabalho é propor uma forma de representar e analisar textos que melhor capture e conserve a maneira como aspectos linguísticos são explorados ao longo dos mesmos e, assim, possa ser utilizada para melhorar a eficácia de tarefas de classificação textual.

Percebendo os resultados promissores dos trabalhos [36; 58; 17; 51; 34], esta pesquisa propõe representar e analisar os textos como fluxos de aspecto (*Aspect Flows*), capturando o comportamento de um aspecto linguístico por toda a extensão do texto.

Os objetivos específicos para a investigação deste trabalho são listados a seguir:

- Adaptar a forma como é realizada a análise de áudio para o domínio dos textos: dado que os textos sejam representados por fluxos que capturam as nuances da variação do comportamento de um aspecto linguístico no texto, assim se assemelhando a sinais de áudio, adequar a forma como a extração de *features* é realizada na análise de áudio para a análise textual;

- Investigar a eficácia proporcionada pela representação de aspectos linguísticos através do uso de léxicos como fluxo de entrada para o método adaptado da análise de áudio: considerando fluxos criados a partir da similaridade semântica textual entre os textos e léxicos relacionados a aspectos alvo da análise, verificar se estes fluxos formam sinais adequados para serem utilizados como entrada do método de análise inspirado no domínio de áudio, permitindo uma boa eficácia do método em tarefas de classificação textual;

- Avaliar a eficácia conferida através do uso das *features* adaptadas das técnicas de análise dos sinais de áudio extraídas dos fluxos em tarefas de classificação de textos;

As questões de pesquisa que este trabalho almeja responder são:

- QP1 - É possível criar um método de análise de textos inspirado na forma como a análise de áudio é realizada, amoldando a extração de *features* para o novo domínio?

- QP2 - É possível obter boa eficácia do método inspirado em análise de áudio em tarefas de classificação textual ao usar, como entrada, fluxos construídos a partir de distâncias semânticas entre texto e léxicos que remetem a um certo aspecto?

- QP3 - O uso de *frames* e *features* adaptadas do domínio de análise de áudio ocasiona boa eficácia em tarefas de classificação de textos?
- QP4 - Quando testada em várias tarefas de classificação de textos, a análise adaptada do domínio de áudio demonstra boa eficácia em todas delas?
- QP5 - O uso de *features* que refletem a variação observada nos fluxos promovem melhores resultados que *features* sintetizadoras em tarefas de classificação textual?
- QP6 - Caso a resposta à questão QP5 seja afirmativa, quão mais eficaz é utilizar *features* que refletem a variação observada nos fluxos quando comparado à utilização de *features* sintetizadoras em tarefas de classificação textual?
- QP7 - O uso de *features* extraídas dos fluxos promovem melhores resultados que o uso dos próprios fluxos em tarefas de classificação textual?
- QP8 - Caso a resposta à questão QP7 seja afirmativa, quão mais eficaz é utilizar *features* extraídas de fluxos quando comparado à utilização dos próprios fluxos em tarefas de classificação textual?

1.2 Contribuições

O trabalho desenvolvido até este momento apresenta algumas contribuições relevantes.

Um método que representa textos através de fluxos que visam acompanhar a forma como aspectos são explorados foi definido. Este método utiliza léxicos e *word embeddings* para capturar os aspectos linguísticos dos textos. Além disso, a análise proposta é inspirada na forma como é realizada a análise de sinais de áudio. Vários experimentos envolvendo diversas tarefas de classificação de texto já foram executados (incluindo diferentes línguas e explorando diferentes aspectos), mostrando que, em algumas situações, esse método tem melhor desempenho que a execução de análise de *features* sintetizadoras, além de revelar mais detalhes sobre como um aspecto é explorado nos textos.

Um artigo descrevendo o método de representação e análise proposto e alguns dos resultados já obtidos foi publicado na conferência intitulada *Language Resources and Evaluation Conference* 2020 - LREC 2020 [56] (Apêndice A).

Até a conclusão do desenvolvimento desta pesquisa de Doutorado, ainda se buscam as seguintes contribuições:

- Usar as *features* decorrentes do método proposto para treinar Redes Neurais Recorrentes, e avaliar o desempenho ao realizar as tarefas de classificação, comparando com os resultados já obtidos;
- Avaliar a eficácia do método proposto ao utilizar modelos de *embeddings* que consideram o contexto das palavras na representação, como ELMo [46] ou BERT [16] para a criação dos *Aspect Flows*, comparando com os resultados já obtidos;
- Aprimorar o método desenvolvido, através da definição e análise de novas *features*;
- Implementar uma solução de tratamento de tamanho dos fluxos criados, como, por exemplo, a normalização de tamanhos de fluxo de um mesmo *dataset*, objetivando facilitar a comparação entre eles; e, ainda, avaliar o impacto gerado nos resultados obtidos nas tarefas de classificação já executadas;
- Avaliar a eficácia do método em tarefas de classificação que envolvam *datasets* contendo textos maiores que aqueles utilizados nos experimentos até então realizados;

1.3 Estrutura do Documento

Esta proposta de tese é estruturada da seguinte forma:

- Capítulo 2: apresenta uma fundamentação teórica sobre conceitos que são base para o desenvolvimento deste trabalho;
- Capítulo 3: apresenta um apanhado de trabalhos relacionados ao desenvolvido nesta proposta;
- Capítulo 4: apresenta o método de representação e análise de textos já proposto;
- Capítulo 5: apresenta os experimentos já realizados e discute dos resultados obtidos;
- Capítulo 6: apresenta as conclusões e o planejamento para a próxima fase do desenvolvimento do trabalho.

Capítulo 2

Fundamentação Teórica

Este capítulo fornece um resumo das informações básicas necessárias para a compreensão desta pesquisa. Inicialmente, o capítulo apresenta uma visão geral da área de Processamento de Linguagem Natural e aborda alguns conceitos relacionados à área que são explorados neste trabalho. Depois, são apresentados conceitos da área de análise de áudio que são importantes para o entendimento do desenvolvimento desta pesquisa.

2.1 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) visa investigar formas de usar computadores para processar e entender a linguagem humana (i.e., natural). PLN é uma área interdisciplinar que envolve linguística computacional, ciência da computação, ciência cognitiva e inteligência artificial. O objetivo da PLN é modelar os mecanismos cognitivos utilizados pelos humanos para entender e produzir sua linguagem. Na prática, PLN se concentra em desenvolver novos modelos e aplicações que facilitem a interação entre as linguagens de máquina e humana. Exemplos de aplicações típicas da PLN incluem: classificação de textos, entendimento de linguagem falada, análise léxica, recuperação de informação, tradução, sumarização de linguagem natural, análise de sentimento, entre outros [15].

Os problemas tratados pelo PLN são muito desafiadores, justamente por lidarem com a linguagem humana. A natureza dessa linguagem é ambígua. Além disso, está sempre mudando e evoluindo. Enquanto os humanos podem facilmente entender a linguagem com toda sua ambiguidade, as regras que regem a passagem de informação através da linguagem

natural não são fáceis de serem entendidas por computadores [23]. Essas regras, muitas vezes, podem ser abstratas e de alto-nível, como o uso de sarcasmo; e outras vezes bem concretas e de baixo-nível como o uso da letra 's' para denotar plural. Toda essa ambiguidade e imprecisão torna difícil o tratamento de linguagem natural por computadores.

2.1.1 *Word Embeddings*

Um dos desafios iniciais na PLN é a forma de representar a linguagem natural. Os modelos baseados em espaço vetorial são uma das formas mais populares de representação. Nesses modelos, cada documento é representado por um vetor cujas dimensões correspondem a *features* extraídas do texto.

Uma forma de representação que se destaca neste contexto é a de representação distribuída, como os *word embeddings*, em que vetores de várias dimensões conseguem exprimir informações do contexto das palavras ao acrescentar à representação uma dependência entre as palavras: quanto mais aproximadas em um contexto, mais dependentes. Em outras palavras, o *word embedding* é uma representação de texto em que palavras que possuem o mesmo significado possuem representação similar [23].

As técnicas mais recentes se utilizam de redes neurais para aprender os *word embeddings* das palavras dos textos através de modelos treinados a partir de grandes massas de dados não tratados. As representações das palavras se dão por vetores densos e de baixa dimensionalidade em um espaço vetorial pré-definido [15]. A representação é capaz de capturar informações sintáticas e semânticas de cada palavra. Goldberg [23] ressalta que o uso de *word embeddings* traz benefícios relativos ao poder de generalização: se algumas *features* são capazes de prover informações similares, é essencial disponibilizar uma representação capaz de capturar essas similaridades.

Exemplos dos chamados *neural embeddings* são o *Word2Vec* [38], *GloVe* [44] e *FastText* [11]. O *Word2Vec* se baseia em duas arquiteturas de redes neurais para produzir uma representação distribuída das palavras:

- *Continuous bag-of-words (CBOW)*: nesta arquitetura, o modelo prediz a palavra mais provável em um dado contexto. Logo, palavras com igual probabilidade de ocorrência no texto são consideradas similares e tomam lugares próximos no espaço vetorial.

- *Skip-gram*: esta arquitetura é similar à CBOW, porém funciona de maneira reversa, ou seja, o modelo prediz o contexto a partir de uma dada palavra.

Diferentemente do *Word2Vec* que se baseia em arquiteturas preditivas considerando apenas o contexto local, o modelo *GloVe* é treinado em uma matriz global de co-ocorrência gerada a partir de uma dada coleção de textos. Essa matriz é decomposta para formar uma representação vetorial mais densa e expressiva. Tanto *Word2Vec* quanto *GloVe* apresentam a desvantagem de não conseguirem representar palavras desconhecidas. Por sua vez, o *FastText* foi proposto para solucionar essa dificuldade. Baseando-se nas arquiteturas do *Word2Vec*, o *FastText* fragmenta as palavras em sub-palavras (*n-grams*) e as alimenta na rede neural. Cada palavra é, então, representada pela soma dos vetores de seus *n-grams*. Uma nova palavra, logo, pode ser representada, uma vez que há grande probabilidade de seus *n-grams* já serem conhecidos pelo modelo treinado.

Uma fragilidade que os modelos *Word2vec, GloVe* e *FastText*, conhecidos como *static word embeddings*, apresentam é a incapacidade de representar a polissemia. Eles retratam cada palavra por um único *word embedding*, independentemente do contexto em que ela ocorre. Para sanar essa dificuldade, recentemente foram propostos modelos que consideram o contexto em que a palavra ocorre para representá-la, os chamados *contextualized word representations*. Logo, a mesma palavra em diferentes contextos possuirá diferentes representações. Como exemplos de *contextualized word representations*, é possível elencar: ELMo [46], BERT [16], and GPT-2 [47].

ELMo, BERT e GPT-2 são modelos de linguagem criados a partir de modelos de *deep learning*. Suas representações internas das palavras são uma função de toda a sentença. ELMo cria representações contextualizadas de cada *token*, concatenando os *internal states* de uma rede *Long Short-Term Memory* (LSTM) bi-direcional de duas camadas treinada sobre modelo de linguagem também bi-direcional [46]. Por sua vez, BERT e GPT-2 são modelos de linguagem baseados em *transformer* [57] bi-direcional e uni-direcional, respectivamente. Cada camada do *tranformer* cria uma representação contextualizada de cada *token* ao participar de diferentes partes da sentença de entrada.

2.1.2 Similaridade Semântica Textual

Similaridade Semântica Textual, ou em inglês *Semantic Textual Similarity* (STS), é o grau de equivalência semântica entre dois textos (ou trechos de textos) [2]. A existência de bons modelos capazes de calcular STS é crucial para muitas aplicações da área de PLN como recuperação da informação, classificação e sumarização de textos [48].

Uma das mais recentes e utilizadas abordagens para o cálculo de STS é a *Word Mover's Distance* (WMD) [31]. Apresentada por Kusner et al., a *Word Mover's Distance* (WMD) é uma função de distância que mede a (dis)similaridade entre dois documentos de texto através da menor distância que os *word embeddings* de um documento precisam "viajar" até alcançar os *word embeddings* de um outro documento. Kusner et al. afirmam que distâncias entre vetores de *word embeddings* são, em certo grau, semanticamente significativas. Utilizando essa propriedade dos *word embeddings*, a WMD representa documentos de texto como uma nuvem de pontos ponderados no espaço vetorial - o conjunto dos *word embeddings* do documento. A distância entre dois documentos é a mínima distância cumulativa que palavras de um documento precisam se "deslocar" até encontrar os pontos da nuvem do outro texto.

A Figura 2.1 ilustra a abordagem por trás do cálculo da WMD. Nela estão representadas duas sentenças provenientes de diferentes documentos nas caixas laterais: (1) "*Obama speaks to the media in Illinois.*" em azul e (2) "*The President greets the press in Chicago.*" em preto. É perceptível que as sentenças, apesar de não possuírem palavras em comum (exceto as *stop words*[1] "*the*" e "*in*"), apresentam praticamente o mesmo sentido. No centro da figura, são ilustradas as palavras que não são *stop words*, representadas no espaço vetorial (*word embeddings*). Os pontos azuis formam a nuvem da sentença (1) e os pretos, da sentença (2). A WMD entre as sentenças (1) e (2) será a soma das mínimas distâncias que cada palavra da sentença (1) precisa "viajar" para encontrar exatamente o ponto correspondente na nuvem do documento ao qual a sentença (2) pertence.

Outras abordagens para cálculo de STS que se pode citar são as que utilizam a *Cosine Similarity* associada a *word embeddings*. A *Cosine Similarity* é uma métrica que determina quão similar dois documentos são, independentemente dos seus tamanhos. Matematicamente, como ilustra a Figura 2.2, é a medida do co-seno do ângulo Θ entre dois vetores pro-

[1] Palavras muito comuns na língua que, em geral, não agregam muito sentido ao texto. São comumente retiradas dos textos antes de serem realizadas as tarefas em PLN.

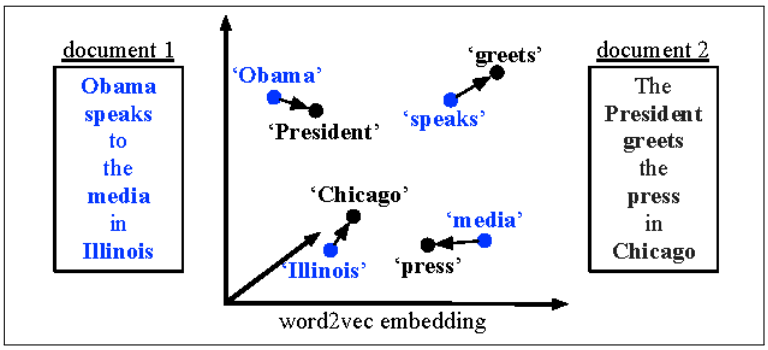

Figura 2.1: Ilutração do *Word Mover's Distance*. (Fonte: Kusner et al. [31])

jetados (V1 e V2) em um espaço multi-dimensional. Se associada a vetores que representam *word embeddings* por métodos como os utilizados por Ranasinghe et al. [48], refletem a STS de dois (trechos de) textos.

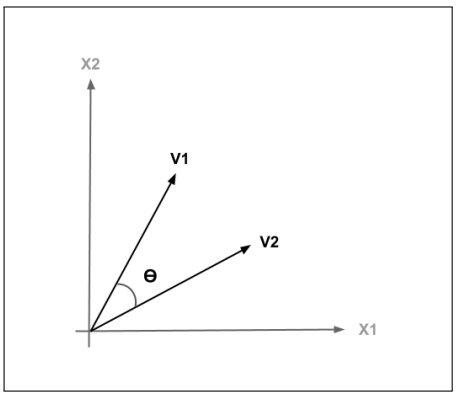

Figura 2.2: Ilustração do ângulo entre dois vetores projetados em espaço bi-dimensional.

2.2 Análise de Áudio

Análise de áudio é uma área de pesquisa bem consolidada que busca estudar e desenvolver conhecimento sobre conteúdo de áudio. O conhecimento já produzido tem se mostrado valioso em vários nichos de tarefas e pesquisa como, por exemplo, segmentação e classificação para recomendação musical [35; 24], classificação de áudio como fala ou música [10], análise de emoção em canções [28] e extração de sentimento em *streams* de discurso [30].

Assim como todo fenômeno natural a ser estudado por meio digital, as ondas sonoras precisam ser representadas como um sinal de tempo discreto. Assim, são utilizadas amostras

das ondas originais para gerar um sinal discreto, sinal este que pode ser representado por vetores de números reais. A Figura 2.3 mostra um gráfico contendo a representação de um sinal sonoro no tempo discreto (*waveform*).

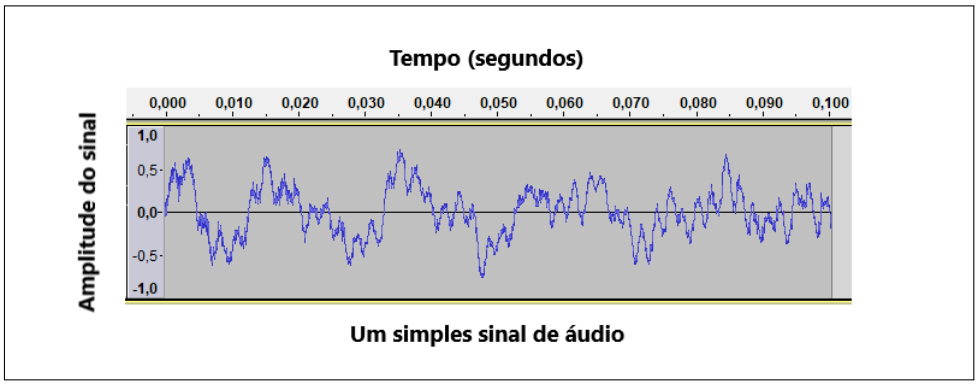

Figura 2.3: Representação de sinal de áudio em tempo discreto. (Adaptada de [20])

Frequentemente, na análise de áudio, o sinal de áudio é fragmentado em partes menores, sobrepostas ou não, chamadas *short-term windows* ou *frames*, e a análise é realizada baseada nesses *frames*. Isto se faz necessário, pois as características de um sinal de áudio, por natureza, variam (muitas vezes rapidamente). Para melhor entender a razão de utilizar essa técnica, considere um áudio que apresenta uma conversa e, no meio dela, acontece um tiro. Se for calculada a intensidade média do sinal considerando todas as amostras, as que apresentam o tiro irão dominar o resultado. Se apenas essa métrica for considerada em uma análise, as conclusões poderão ser deturpadas. Realizar análises usando os *frames* permitem capturas de informações mais fidedignas ao conteúdo do áudio [19].

Uma vez obtida a divisão do sinal de áudio em *frames*, é possível realizar a extração de *features* que formam uma valiosa representação do sinal, identificando caraterísticas importantes do mesmo e, portanto, sendo bem úteis para posterior análise. Essas *features* podem ser classificadas quanto ao tipo e quanto ao domínio. Quanto ao tipo, elas podem ser chamadas de *short-term features*, quando extraídas de cada *frame* individualmente, remetendo a características do próprio *frame*; ou *mid-term features*, quando estatísticas, como média, variância, desvio padrão, são calculadas a partir de *short-term features* de *frames* consecutivos que compõem um *mid-term segment* ou *window*. Por sua vez, quanto ao domínio, as *features* podem ser do domínio do tempo, quando extraídas diretamente da representação do sinal de áudio a partir de amostras em tempo discreto; ou do domínio da frequência (ou espectro,

como também é conhecido), quando são extraídas a partir da representação da distribuição de frequência do conteúdo do sinal sonoro (espectro do som) [22]. Esta representação é obtida através da aplicação da Transformada Discreta de *Fourrier* (TDF) à representação do sinal de áudio em tempo discreto [21].

A literatura apresenta uma grande variedade de *short-term features* tanto do domínio do tempo quanto da frequência [22; 27; 10]. Neste capítulo serão abordadas as *features* que serviram de inspiração para a criação das *features* utilizadas nesta pesquisa e, portanto, são necessárias para o entendimento da mesma.

2.2.1 *Features* do Domínio do Tempo

As três *features* do domínio do tempo que serão abordadas neste capítulo são: *Energy*, *Zero-Crossing Rate* e *Energy Entropy*.

Energy

A *feauture Energy* reflete a magnitude da amplitude do sinal de áudio [27].

Seja $x_i(n), n = 1, ..., W_L$ a sequência de amostras de áudio do i-ésimo *frame*, onde W_L é o comprimento do *frame*. A implementação da *Energy* é definida segundo a Equação 2.1:

$$E(i) = \frac{1}{W_L} \sum_{n=1}^{W_L} |x_i(n)|^2 \tag{2.1}$$

Na Equação 2.1, a Energy foi dividida por W_L, normalização necessária para remover a dependência do tamanho do *frame*.

Zero-Crossing Rate

Zero-Crossing Rate (ZCR) é a taxa de mudanças de sinal dentro do *frame*. Ou seja, é o número de vezes que o sinal muda seu valor, de positivo para negativo e vice-versa, dividido pelo comprimento do *frame* [22].

A *ZCR* é definida pela Equação 2.2:

$$ZCR(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{2.2}$$

onde sgn é a função sinal, denotada pela Equação 2.3:

$$sgn[x_i(n)] = \begin{cases} 1, & \text{se } x_i(n) \geq 0. \\ -1, & \text{se } x_i(n) < 0. \end{cases} \tag{2.3}$$

A *ZCR* pode ser interpretada como a medida de ruído de um sinal de áudio. Esta *feature* é conhecida por refletir, de maneira mais simplificada, as características espectrais do sinal.

Entropy of Energy

A *feature Entropy of Energy* pode ser interpretada como uma medida de mudanças abruptas no nível de energia de um sinal de áudio [22].

Primeiramente, para computá-la, é necessário dividir cada *frame* em K *sub-frames* de duração fixa. Daí, é computada a *Energy* de cada *sub-frame* j usando a Equação 2.1 que é, então, dividida pela *Energy* total do *frame*, $Eframe_i$. Essa divisão é necessária para que se possa tratar a sequência valores de *Energy* dos *sub-frames*, e_j, $j = 1, ..., K$, como uma sequência de probabilidades, como na Equação 2.4:

$$e_j = \frac{Esubframe_j}{Eframe_i} \tag{2.4}$$

onde

$$Eframe_i = \sum_{k=1}^{K} Esubframe_k \tag{2.5}$$

Ao final, a *Entropy of Energy*, $H(i)$ da sequência e_j é computada segundo a Equação 2.6:

$$H(i) = -\sum_{j=1}^{K} e_j * log_2(e_j) \tag{2.6}$$

É necessário salientar que, quanto mais mudanças significativas presentes na energia do *frame*, menor é o valor da *Entropy of Energy*.

2.2.2 *Features* do Domínio da Frequência

Para computar as *features* do domínio da frequência, primeiro é necessário computar a TDF dos *frames* do áudio e usar os coeficientes resultantes como entrada para o cálculo das *fea-*

tures.

Para o entendimento das *features* que serão discutidas, considere $X_i(k), k = 1, ..., Wf_L$ como sendo a magnitude dos coeficientes do i-ésimo *frame* resutante da TDF.

Das muitas *features* do domínio da frequência existentes, esta seção se atém a duas que serviram de inspiração no trabalho: *Spectral Entropy* e *Spectral Flux*.

Spectral Entropy

Spectral Entropy é computada de maneira similar à *Energy Entropy*, porém, agora no domínio da frequência. Mais especificamente, primeiro é necessário dividir o espectro do *frame* em L *sub-bands* (ou *bins*). A energia E_f da f-ésima *sub-band*, $f = 1, ..., L - 1$, é então normalizada pela energia espectral total do *frame*, dada pela Equação 2.7:

$$n_f = \frac{E_f}{\sum_{f=0}^{L-1} E_f}, f = 1, ..., L - 1 \tag{2.7}$$

A entropia do espectro de energia normalizado n_f é então computada segundo a Equação 2.8:

$$H(i) = -\sum_{f=0}^{L-1} n_f * log_2(n_f) \tag{2.8}$$

Spectral Flux

Spectral Flux mede as mudanças que ocorrem no espectro considerando dois *frames* sucessivos. É computada pelo quadrado da diferença entre as magnitudes de espectro normalizadas desses dois *frames*, como mostra a Equação 2.9:

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} (EN_i(k) - EN_{i-1}(k))^2 \tag{2.9}$$

onde $EN_i(k)$ é o k-ésimo coeficiente da TDF do i=ésimo *frame*, calculado como mostra a Equação 2.10.

$$EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} (X_i(l)} \tag{2.10}$$

2.3 Considerações Finais

Neste capítulo, foram abordados, além da definição da área de PLN, os conceitos de *Word Embeddings* e Similaridade Semântica Textual, tópicos da área de PLN utilizados no decorrer desta pesquisa e que podem não ser de amplo conhecimento do leitor desse documento. Também foram abordados conceitos da área de análise de áudio, área em que o método proposto neste trabalho se inspira e, portanto, necessários para que seja entendida parte da pesquisa. No próximo capítulo, serão discutidas pesquisas que se dedicaram em manter a forma como as informações são exploradas ao longo do texto em suas representações. Também serão discutidos trabalhos que visam capturar e trabalhar com informações semânticas extraídas do texto.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentados os esforços de alguns estudos anteriores em representar e analisar textos de forma a captar informações que contemplem o conteúdo ao longo dos textos, sem se utilizar de generalizações para resumir informações dos textos de uma forma global na Seção 3.1. Por sua vez, a Seção 3.2 apresenta alguns estudos que focam em representar os textos de forma a captar informações semânticas neles presentes.

3.1 Representação e Análise Contemplando Conteúdo ao Longo dos Textos

Uma abordagem para evitar representar um texto de forma sintetizada e global é modelar o mesmo como um fluxo, que pode ser definido como uma sequência de informações coletadas a partir de palavras, sentenças ou parágrafos de um texto [36]. Tendo o fluxo como representação, a análise deve ser feita respeitando as características dos fluxos, mantendo a sequência das informações e evitando gerar métricas que apenas resumam todas as informações da sequência, como por exemplo médias e medianas. Manter as características do fluxo pode permitir uma análise mais aprofundada da forma como os textos foram escritos.

Em seu estudo, Mao e Lebanon [36] usam fluxos de sentimentos para representar os textos. os autores propõem uma variação do método de modelagem estatística *Conditional Random Fields (CRF)* [33], o *Isotonic Conditional Random Fields (ICRF)*, que contempla dados ordinais para predizer o sentimento de cada sentença, denominado de sentimento lo-

cal. A sequência de sentimentos locais preditos forma o fluxo de sentimentos de um texto. Também verificam que fluxos de diferentes textos terão diferentes tamanhos, o que tornaria difícil a comparação entre eles ou o treinamento de modelos que os utilizassem. Para solucionar este problema, eles propõem uma normalização dos fluxos dos textos, utilizando uma função de suavização dos valores das sentenças, transformando o fluxo em uma curva de transições mais suaves. Isto permite o cálculo de distâncias entre elas e, ainda assim, mantém as mudanças de sentimentos entre as sentenças.

Os experimentos apresentados no trabalho [36] mostram que o *ICRF* obtém melhores resultados que o *CRF*, *Naive-Bayes* e SVM ao predizer os sentimentos locais. Mostram ainda que ao utilizar fluxos de sentimentos, conseguem predizer melhor os sentimentos globais dos textos que uma representação por *bag-of-words (BOW)*, já despertando o poder da representação por fluxos. Mostram ainda que a análise da forma como se dá mudança de sentimentos entre sentenças dos fluxos pode denotar diferenças de estilos de escrita de autores distintos. Além disso, ainda ressaltam o potencial uso da representação de textos através de fluxos de sentimento na tarefa de sumarização.

Por outro lado, Wachsmuth e Stein [58] apresentam a representação da estrutura de discurso de um texto como fluxos de *rhetorical moves* (funções comunicativas de segmentos de textos argumentativos, geralmente ligadas ao objetivo final da argumentação). Eles modelam até quatro tipos de fluxos de texto, a saber: sentimento local, modelando sentimentos negativos, positivos e neutros; relação de discurso entre segmentos de texto, por exemplo: causa, circunstância ou condição; funções de discurso em nível de parágrafo, por exemplo: introdução, corpo, conclusão; e papéis da argumentação, modelando os segmentos como argumentos reais, premissas ou afirmações.

O objetivo do trabalho de Wachsmuth e Stein [58] é capturar a estrutura global de um texto argumentativo comparando o seu fluxo com um padrão de fluxos comum extraído de um conjunto de textos de treinamento. Para a realização dessa comparação, é necessária a unificação dos fluxos para a extração de um padrão. São sugeridas duas formas de unificação: normalização dos fluxos para um mesmo espaço vetorial e abstração de variações entre fluxos, mapeando fluxos similares para um mesmo fluxo. Para a normalização, é determinado um tamanho alvo de texto, e usada interpolação nos fluxos para que todos se adequem ao tamanho escolhido. Ao tratar do método de abstração, os autores sugerem excluir três

comportamentos: *rhetorical moves* iguais em sequência, deixando apenas um como representante; ciclos de dois ou mais *rhetorical moves* idênticos (deixando apenas um ciclo); e exclusão de classes de *rhetorical moves* de menor importância. Após unificar os fluxos, é realizada uma clusterização dos fluxos de treinamento, usando distâncias de similaridade em espaço vetorial como *Manhattan Distance* para os fluxos normalizados e *Minimum Edit Distance* para fluxos abstraídos. Daí, para verificar a qual *cluster* de fluxos um fluxo de teste é mais similar, usam as mesmas distâncias entre o fluxo de teste e o *centroid* dos *clusters* de treinamento.

Os resultados de alguns dos experimentos apresentados por Wachsmuth e Stein [58] mostram que vários padrões de fluxo podem ser encontrados nos *datasets* de avaliações de produtos e serviços e de redações trabalhados e que alguns *rhetorical moves* são mais presentes que outros. Ainda, que usar abstrações pode deixar o modelo robusto a mudanças de domínio.

Em se tratando dos experimentos de classificação de sentimento global, o trabalho de Wachsmuth e Stein [58] apresentou avaliações de produtos, hotéis e filmes, variando entre classificações dentro do mesmo domínio e entre domínios diferentes e ainda usando dois tipos de obtenção dos sentimentos locais: o *ground-truth* dos *datasets* e a classificação utilizando o método de Socher et al. [54]. Os resultados mostram que em alguns casos os fluxos classificam melhor os sentimentos globais; em outros, os métodos baseline (*BOW* e frequência de sentimentos) obtêm melhores marcas; e, ainda, uma combinação entre fluxos e *baselines* obtém melhor acurácia em outros casos. Ou seja, não há apenas um método que seja melhor em todas as situações testadas. O trabalho também apresentou experimentos de avaliação de organização de redações, tendo mostrado que usando as informações das sentenças dos fluxos eles obtiveram melhores resultados que o método estado da arte referenciado (Persing et al. [45]).

Em sua pesquisa, Filatova [17] estuda a detecção de sarcasmo em textos grandes como avaliações de produtos. Ela mostra que características que denunciam sarcasmo em textos pequenos mensagens de mídias sociais como emoticons, hashtags e pontuação forte não tem o mesmo efeito em textos maiores. Em textos grandes, ela afirma que entender o contexto é muito importante, pois o sarcasmo é detectado em frases que tem uma certa polarização, em um contexto contrário. Sendo assim, o trabalho propõe modelar as avaliações de produtos como fluxos de sentimentos como forma de capturar o contexto da avaliação e usa a mudança

de sentimentos entre sentenças (de negativo para positivo e vice-versa) para a detecção de sarcasmo.

Os resultados dos experimentos executados mostram que o método proposto por Filatova [17] obtém melhores resultados que o *random baseline* quando são consideradas todas as avaliações e, principalmente, quando consideradas apenas as avaliações positivas. Em se tratando das avaliações negativas, o método apresenta resultados bem próximos aos do *random baseline*, o que sugere que o comportamento do sarcasmo nessas avalições difere do comportamento nas avaliações positivas.

No domínio de recuperação de documentos relevantes, Seo e Jeon [51] alegam que os modelos tradicionais de *BOW* usam estatísticas para medir a relevância dos documentos, o que dificulta a detecção de documentos não-relevantes que contenham vários termos da consulta de forma aleatória ou estejam fora do contexto. Como solução para este problema, os autores propõem uma representação dos textos através de um fluxo de relevância, em que é calculado o nível de relevância de cada sentença em relação à consulta.

Na implementação da solução proposta por Seo e Jeon [51], são gerados os fluxos de relevância dos N documentos melhor ranqueados pelo motor de busca *baseline* e são extraídas as seguintes *features* desses fluxos: média e variância do nível de relevância; razão entre número de picos e número de sentenças do documento; posição do primeiro pico; e média e variância das posições dos picos no documento. Os resultados mostram que há uma melhora estatisticamente significativa quando se faz o novo ranqueamento dos documentos utilizando as *features* extraídas dos fluxos.

De forma similar a Seo e Jeon [51], a investigação de Lee et al. [34] representa documentos como um fluxo de sentimento-relevância para buscar melhores resultados de consultas a documentos que contenham opinião. Para cada sentença dos documentos, calculam uma pontuação que reflete a relevância em relação a uma consulta e a opinião dada pela frequência da aparição de palavras de um léxico. As *features* extraídas são a variância entre os valores das sentenças, razão entre número de picos e número de sentenças do documento e a posição do primeiro pico. Os experimentos seguem o mesmo formato do trabalho de Seo e Jeon [51] e seus resultados mostram que a abordagem usando fluxos melhoram significativamente a qualidade dos resultados das buscas.

3.2 Representação de Textos Sob uma Perspectiva Semântica

A forma de representar a linguagem natural dos textos tem sido um desafio para a área de PLN, sendo os modelos de espaço vetorial uma das formas mais populares de representação, em que cada documento é representado por um vetor cujas dimensões correspondem a *features* extraídas do texto. Um exemplo desse tipo de modelo é o *BOW*, em que as *features* são palavras independentes. Neste tipo de modelo, não há como capturar informações semânticas, mas mesmo simples, é um modelo que obtém bons resultados em várias tarefas, sendo até difícil de suplantá-lo em diversos cenários.

Ainda assim, há cenários em que se faz necessária uma representação mais rica em termos de semântica, para que se possa descobrir padrões mais elaborados e, assim, estudar mais a fundo a forma como humanos se expressam através de textos [1]. Uma forma de representação que se destaca neste contexto é a de representação distribuída, como os *word embeddings*, em que vetores de várias dimensões conseguem exprimir informações do contexto das palavras ao acrescentar à representação uma dependência entre as palavras: quanto mais aproximadas em um contexto, mais dependentes. O uso de *word embeddings* vem se tornando bem popular e se provando uma forma de representação que consegue apresentar muitas informações significativas, como mostra o trabalho de Baroni, Dino e Kruszewski [9]. Ao comparar o desempenho dos *word embeddings* frente a modelos baseados em contagem de co-ocorrência em diferentes tarefas semânticas, os autores chegam à conclusão de que os modelos baseados em *word embeddings* obtiveram melhores resultados.

Uma outra tendência de pesquisa nessa área é propor modelos que, baseados em *word embeddings*, ainda acrescentam mais conhecimento à representação. Um exemplo deste tipo de trabalho é o de Sinoara et at. [53] que apresenta duas abordagens para representação semântica de coleções de documentos, a NASARI+Babel2Vec e a Babel2Vec, baseadas na eliminação da ambiguidade usando *word senses* e *embeddings* baseados em palavras e/ou *word senses*.

As representações propostas na pesquisa de Sinoara et at. [53] se utilizam do mesmo espaço vetorial que os *embeddings* e não precisam de grandes quantidades de documentos para treinar os modelos. Além disso, podem ser utilizadas em tarefas de classificação de

várias classes. O método NASARI [12], usado como parte de uma das abordagens, constrói vetores de *word senses* que podem ser aplicados a várias línguas, pois é baseado no BabelNet [40]. Ao analisarem os vetores que representam os documentos, os autores afirmam que eles indicam que ambas as representações apresentam vetores próximos a palavras ou *word senses* relacionadas, mas a representação do NASARI+Babel2Vec é mais vantajosa, pois a vizinhança baseada em *word senses* tem mais significado e é mais facilmente interpretável.

Para avaliar o método proposto, Sinoara et at. [53] detalham vários experimentos realizados utilizando seis algoritmos de aprendizagem de máquina, nove *datasets*, vários tipos de tarefas de classificação em diferentes níveis de dificuldade e duas línguas: inglês e português. Os resultados destes experimentos indicam forte desempenho por parte das abordagens propostas, especialmente nos cenários mais complexos dos *datasets* em língua inglesa. Entretanto, nos experimentos em língua portuguesa, as representações usando *BOW* obtiveram melhores resultados, ainda que não tenham sido considerados bons. Os autores alegam que os maus resultados das abordagens propostas na classificação de língua portuguesa se devem à pequena cobertura dos recursos linguísticos nesta língua, o que leva a acreditar que há espaço para melhorias.

Uma outra forma de adicionar informações relevantes à representação de um texto é fazer uso de léxicos. Léxicos, no domínio do PLN, são conjuntos de palavras e/ou expressões que refletem um certo aspecto linguístico de uma língua. Ou seja, a presença de um ou mais componentes de um léxico em sentenças ou documentos garante que o aspecto linguístico representado por ele está presente naquele texto. Por exemplo, se um texto apresenta palavras presentes em um léxico de sentimento ou de argumentação, indica que o autor do texto quis expressar sentimento ou quis se utilizar de ferramentas de argumentação daquela língua no texto.

Uma tendência na busca de melhores representações de texto é associar o poder da representação distribuída dos *word embeddings* com o conhecimento adicional que os léxicos promovem. Isto porque os *word embeddings* pré-treinados (amplamente utilizados) contém informações semânticas e sintáticas para um contexto geral. É possível aliar tudo isso ao conhecimento sobre um aspecto específico, ao incluir as informações de um léxico abordando o aspecto em questão.

Muitos trabalhos exploram essa tendência através da similaridade semântica, ou seja, do

cálculo de distâncias semânticas entre os vetores de representação das palavras de sentenças e documentos geradas através dos *word embeddings* e as palavras dos léxicos. Assim, é possível representar o texto capturando o aspecto linguístico representado pelo léxico de uma forma mais completa, ainda que os textos não apresentem exatamente as palavras que formam o léxico.

Araque et al. [6] conduziram uma pesquisa em que propõem um modelo de análise de sentimentos que explora justamente a distância semântica entre *word embeddings* e léxicos. O modelo sugerido gera um vetor de features para cada documento que é fruto da concatenação de dois outros vetores: o primeiro, um vetor de representação do texto por *word* ou *document embeddings*; o segundo, um vetor de similaridade entre palavras do texto e um léxico de sentimento.

Para o cálculo do vetor de similaridade, primeiramente, o modelo de Araque et al. [6] escolhe palavras do léxico que são mais frequentes no conjunto de treinamento. Daí, gera uma matriz de similaridade calculando o produto escalar entre as palavras do texto e as palavras escolhidas do léxico e, só então, reduze essa matriz a um vetor, escolhendo o maior valor de similaridade obtido para cada palavra do léxico.

Os autores, ainda na pesquisa [6], apresentam resultados de vários experimentos conduzidos utilizando *datasets* de textos curtos e longos e ainda uma variedade de léxicos. Reportam os resultados da classificação usando apenas o vetor de representação por *word embeddings*, apenas o vetor de representação por similaridade semântica e a representação do modelo proposto que concatena os dois tipos vetores. Os resultados mostram que o método proposto apresenta melhor desempenho em relação à classificação apenas por similaridade semântica em todas as combinações de *dataset* e léxico. Fato que não se repete quando comparado à classificação apenas utilizado os *word embeddings*.

Por sua vez, a investição de Jerônimo et al. [29] propõe uma representação de textos baseada em subjetividade para realizar a tarefa de classificação de notícias falsas. Apostando que os níveis de subjetividade de notícias legítimas e falsas são significativamente diferentes, os autores se utilizam de cinco léxicos de subjetividade em português do Brasil criados por linguistas brasileiros [4] para criar vetores de subjetividade como *features* para cada documento de um *dataset* de notícias brasileiras.

Jerônimo et al. [29] criam os vetores de subjetividade através do cálculo da distância

WMD [31] entre cada uma das sentenças das notícias e os léxicos de subjetividade considerando os *word embeddings* das palavras presentes nas notícias. Então, uma média das distâncias de todas as sentenças de cada documento para cada léxico é calculada, sendo o vetor de features formado, finalmente, por essas cinco médias.

Os resultados dos experimentos apresentados no trabalho [29] se mostram promissores, pois, apesar de no experimento que classifica as notícias em geral, o método proposto se mostra tão bom quanto o TF-IDF (baseline), o método proposto obtém melhores e mais robustos resultados quando cenários entre domínios (tipo de notícias - cultura, esporte, política e economia - ou fontes - Estadão e Folha) são considerados.

Outra forma que tem sido explorada é adicionar o conhecimento agregado dos léxicos a redes neurais, para, principalmente, obter melhores resultados na tarefa de classificação de sentimentos.

Nessa vertente, há trabalhos como o de Wu et al. [61] e o de Fu et al. [18] que incluem informações do léxico nos *word embeddings* que representarão os textos. No primeiro trabalho, os autores usam o léxico para treinar um classificador de sentimentos de palavras e usa esse classificador para criar um *word sentiment embedding*, que é concatenado com o word embedding original pra representar as palavras dos textos e alimentar a rede neural para classificação de sentimentos. Exaustivos experimentos mostram que o modelo proposto obtém melhores resultados que outros modelos usados como baselines.

Já no segundo trabalho [18], os autores criam um mecanismo de atenção baseado na correlação dos vetores dos *word embeddings* das palavras de cada sentença com os vetores dos *word embeddinngs* das palavras do léxico (positivo e negativo), correlação esta, calculada pelo produto escalar. A representação final das sentenças resolve o problema de ambiguidade semântica das palavras e alimenta uma rede LSTM bidirecional. A saída dessa rede alimenta um outro mecanismo de atenção que captura informações importantes sobre o contexto de diferentes subespaços de representação em diferentes posições. Só então, os dados passam pela camada de classificação. Dos experimentos em quatro *datasets* de classificação de sentimentos, o modelo proposto neste trabalho obtém melhores resultados que outros métodos baseline em três deles.

3.3 Considerações Finais

Neste capítulo, foram apresentados alguns estudos que, ao representar textos através de fluxos que capturam informações ao longo dos textos, obtêm bons resultados nos experimentos realizados, demonstrando alguns dos benefícios que esse tipo de representação pode trazer. Também através da discussão de alguns trabalhos, foram ressaltados os benefícios propiciados por representações que consideram características semânticas dos textos. Essa discussão se deu necessária, pois se relaciona com a pesquisa apresentada neste documento, uma vez que esta visa representar e analisar textos via aspectos linguísticos - que remetem a características semânticas do texto - de maneira a manter a forma como o aspecto é explorado ao longo de todo o texto. O método proposto para alcançar os objetivos desta pesquisa será apresentado no próximo capítulo.

Capítulo 4

Método Proposto

Neste capítulo, será apresentado o método proposto para representação e análise de textos, que visa captar aspectos linguísticos com ênfase em manter a forma como são explorados ao longo de todo o texto. Objetivando, assim, um melhor entendimento de como um dado aspecto é explorado em determinado tipo de texto, ao extrair *features* que refletem o comportamento do aspecto ao longo dos textos. Essas *features* são utilizadas para alimentar algoritmos de *Machine Learning* a fim de realizar tarefas de classificação.

Diante dos bons resultados obtidos pelos trabalhos que usam fluxos como representação de textos [36; 58; 17; 51; 34], o método apresentado neste trabalho propõe representar os textos como fluxos de aspecto (*Aspect Flows*), capturando as informações semânticas acerca de um aspecto linguístico por toda a extensão do texto. Após essa etapa, visando uma forma de entender como um aspecto se comporta no texto, a abordagem se inspira na área de análise de áudio para realizar uma sofisticada análise dos *Aspect Flows* baseada no conceito de *frames*. Dessa forma, é possível extrair *features* que detêm informações relevantes sobre o texto, mantendo informações individualizadas de cada parte (*frame*), o que pode levar a um melhor conhecimento sobre o comportamento do aspecto. No decorrer do capítulo, o método proposto será descrito em detalhes, compreendendo a criação de *Aspect Flows*, a divisão em *frames* e a subsequente extração das chamadas *Audio-Like Features*, adaptações de algumas das *features* presentes na análise de áudio para o domínio de análise de textos.

A Figura 4.1 apresenta um diagrama do método proposto. Como é mostrado, o método cria a representação do aspecto explorado no texto como um *Aspect Flow*. Em seguida, o *Aspect Flow* é fragmentado em um certo número de *frames* e, daí, a partir de cada *frame*,

são extraídas as *Audio-Like Features*. Essas *features* são, então, utilizadas para alimentar algoritmos de *Machine Learning* para realizar tarefas de classificação.

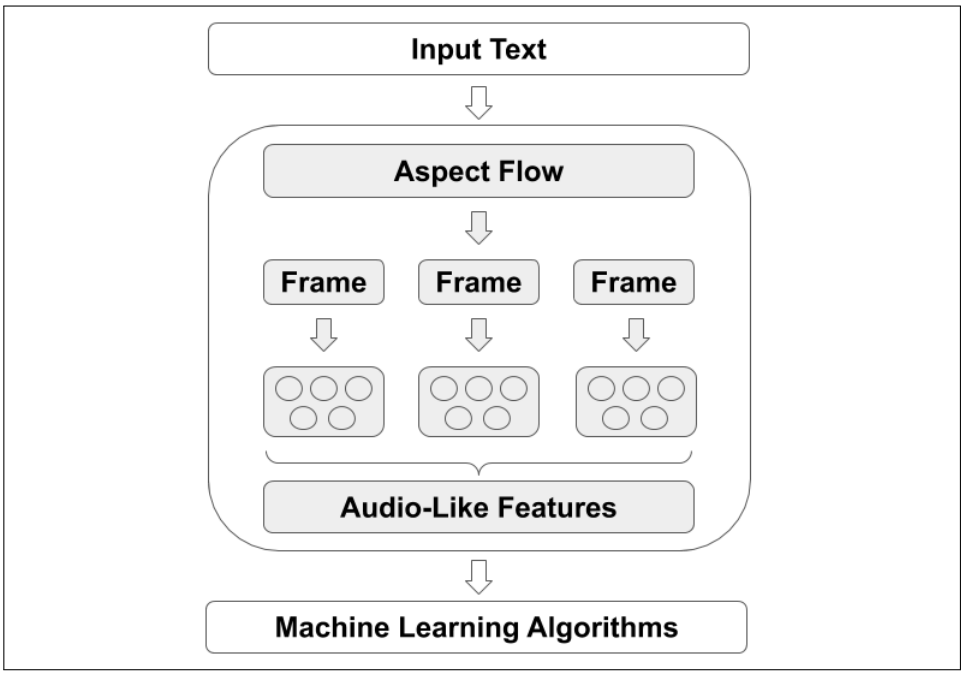

Figura 4.1: Método Proposto

4.1 Criação dos *Aspect Flows*

Representar textos por um fluxo de um aspecto relacionado a uma tarefa de PLN é uma maneira promissora de melhor entender como um aspecto se comporta no texto, como a modelagem de fluxos de sentimento explorada no trabalho de Mao e Lebanon [36] confirma.

Para gerar fluxos de aspectos, é necessário que os textos sejam divididos em unidades menores, como parágrafos, sentenças ou até palavras, e, então, obter uma representação que remeta ao aspecto para cada uma dessas unidades. O fluxo do texto será formado pela sequência de todas essas representações. Este método utiliza sentenças como unidades para geração de fluxos.

Há várias formas de obter as representações dos aspectos para cada sentença. Uma delas é utilizar uma base de dados anotada por sentença compreendendo textos e aspecto relacionados a uma determinada tarefa. Essa opção é bem limitada, pois há poucas bases que atendam a esse requisito disponíveis, além de serem atreladas a um único aspecto, o mais

comumente disponível sendo sentimento, e a uma língua. Criar bases anotadas manualmente é um trabalho muito custoso, o que limita as opções de aspectos e tarefas.

Uma outra opção é seguir os passos de Filatova [17] e usar anotadores automáticos como o Standford Sentiment Analysis Tool (SSAT) [55], que permite anotar as sentenças em uma escala de sentimentos de 5 pontos (muito negativo (-2), negativo (-1), neutro (0), positivo (1), muito positivo (2)). A Figura 4.2 apresenta um *Aspect Flow* criado a partir de anotações SSAT de um trecho de avaliação de filme. O trecho apresenta cinco sentenças, cada uma recebe uma anotação SSAT e a sequência dessas anotações formam o *Aspect Flow*. Essa alternativa evita a necessidade de anotação manual de bases, mas ainda é limitada em relação aos tipos de aspectos que podem ser explorados e as línguas a serem trabalhadas.

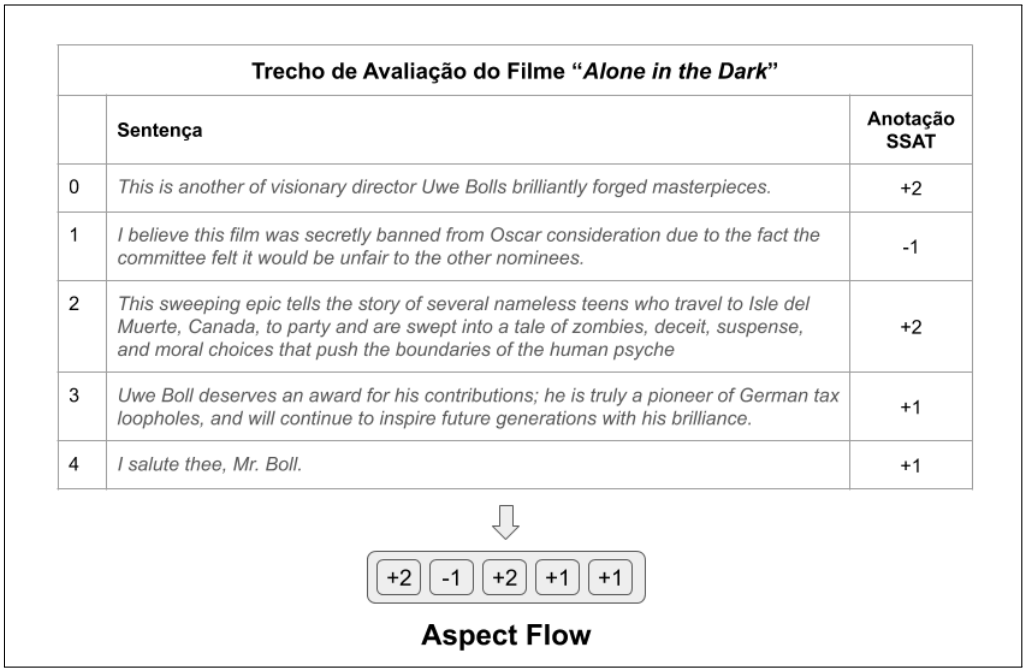

Figura 4.2: Exemplo de um *Aspect Flow* criado a partir de anotações SSAT de um trecho de avaliação de filme (trecho extraído do artigo de Filatova [17])

Uma alternativa para criação dos *Aspect Flows* mais flexível quanto a língua e aos aspectos a serem trabalhados é o cálculo de similaridade semântica entre as sentenças do texto e um léxico de palavras que reflita o aspecto em uma determinada língua. Neste caso, pode-se usar um modelo de *word embeddings* na língua alvo da tarefa, e então calcular uma métrica de similaridade semântica textual (como WMD) entre as sentenças e o léxico no espaço de *embedding*. A Figura 4.3 ilustra a criação de um *Aspect Flow* genérico: é calculada a distân-

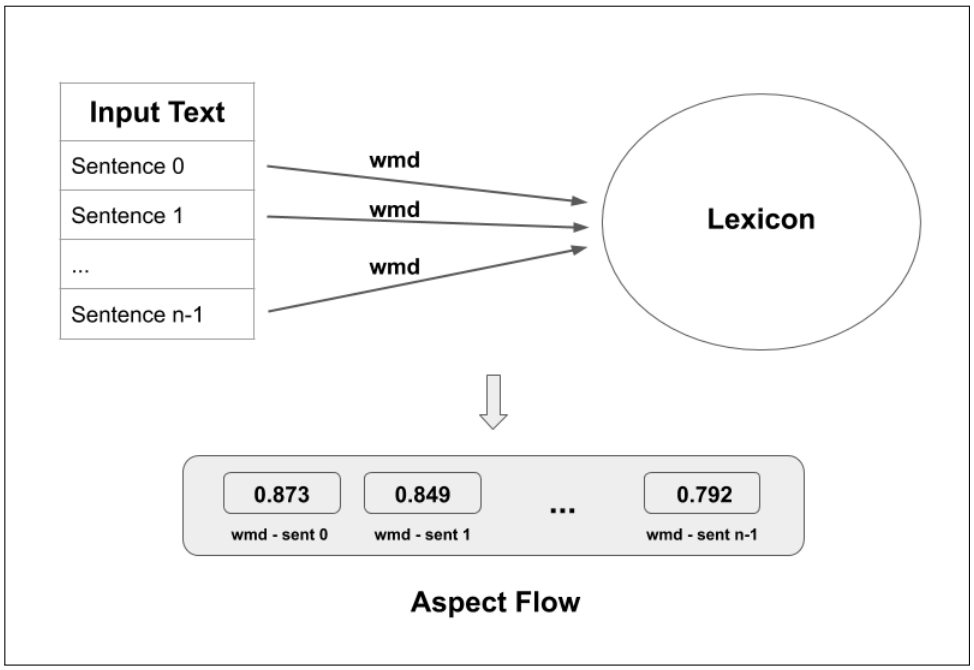

Figura 4.3: Ilustração de um *Aspect Flow* genérico criado a partir de distâncias WMD das sentenças de um texto a um léxico.

cia WMD de cada sentença do texto ao léxico (ambos representados por *word embeddings*) e a sequência dessas distâncias forma o *Aspect Flow*. As distâncias, neste caso, serão valores pertencentes ao intervalo [0,1], quanto menores, maior a similaridade apresentada entre as sentenças e o léxico.

Nesta opção, o método dependerá do treinamento de modelo de *word embeddings* e de léxicos relacionados a um aspecto em determinada língua. No tocante aos modelos de *word embeddings*, há alguns modelos pré-treinados amplamente utilizados, principalmente em inglês; e, ainda, é possível treinar modelos sem enorme esforço. Quanto aos léxicos, há vários disponíveis na literatura tratando, por exemplo, de sentimento, de subjetividade, de argumentação, novamente, principalmente em inglês; mas também há a possibilidade da criação de novos léxicos ou da tradução de léxicos já existentes para outras línguas. Ao usar essa alternativa, o método pode ser independente de língua e de aspecto, se tornando mais adaptável a várias tarefas.

4.2 Divisão dos *Aspect Flows* em *Frames*

Se for plotado um *Aspect Flow* em um gráfico usando o eixo-x para representar as sentenças de um texto e o eixo y os valores que representam o aspecto, é possível perceber similaridades entre a forma desse gráfico e a de um gráfico de sinal de áudio em tempo discreto (como no exemplo da Figura 2.3). Para ilustrar essas semelhanças, guardadas as devidas proporções em relação ao número de amostras que é bem maior no sinal de áudio, a Figura 4.4 mostra um exemplo de *Aspect Flow* de subjetividade de um texto. O gráficoc mostra os níveis de subjetividade associados às sentenças ao longo do texto.

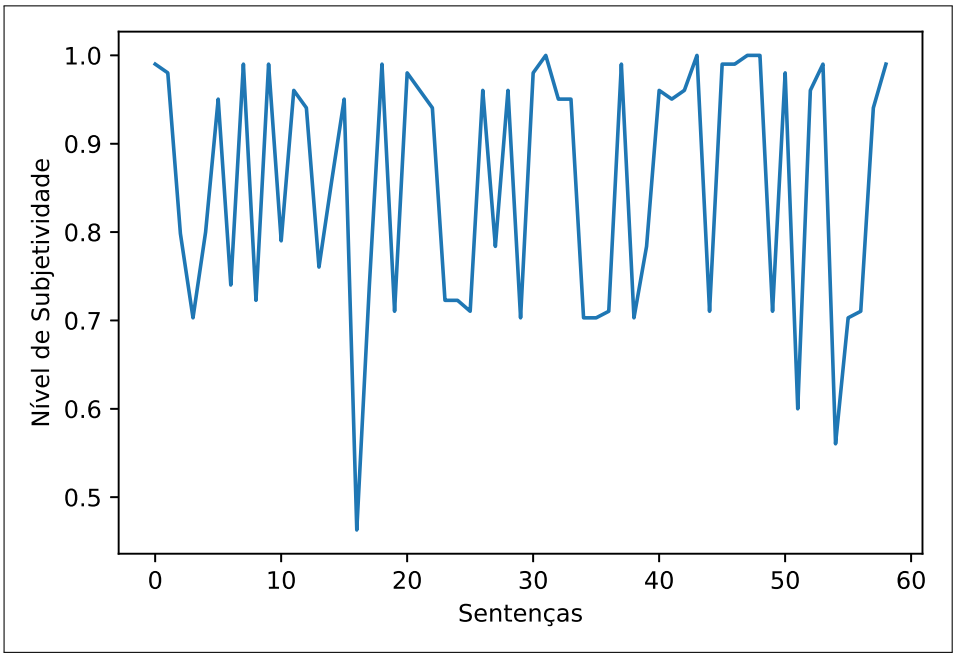

Figura 4.4: Exemplo de Gráfico de um *Aspect Flow* de Argumentação

Inspirada por essa similaridade, este método propõe adaptar a maneira como a análise de áudio é realizada para o domínio do PLN.

No Capítulo 2, foi explicado o motivo do uso da divisão em *frames* na análise de áudio que leva a computação de melhores *features* para representar todo o conteúdo de um arquivo de áudio.

Como a ideia deste trabalho é examinar como os aspectos são explorados ao longo de todo o texto, a proposta do método é adotar a técnica de *short-term windowing*, fragmentando os *Aspect Flows* em *frames*, inicialmente não sobrepostos.

Para ser capaz de comparar os mesmos trechos de diferentes textos (que muito frequentemente possuem tamanhos diferentes), o método aqui proposto fragmenta os *Aspect Flows* em um número fixo de *frames*, diferentemente do método de fragmentação de análise de áudio que leva o tamanho do *frame* em consideração. Assim, independentemente do número de sentenças em um texto, o primeiro *frame* representará a primeira parte do texto, por exemplo, que poderá então ser comparada com a primeira parte de outro texto. A definição do número de *frames* a fragmentar os fluxos depende do *dataset* trabalhado, uma vez que, por exemplo, ao tratar com livros é possível obter muito mais *frames* que ao tratar com avaliações de produtos ou serviços.

4.3 Extração de *Audio-Like Features* a partir dos *Aspect Flows*

Na análise de áudio há dois tipos de *features* extraídas dos *frames* dos arquivos de áudios em relação ao domínio: *features* do domínio do tempo e da frequência. Este método inicialmente se propôs a adaptar a extração de algumas *features* do domínio do tempo, por ser possível realizá-la diretamente dos *Aspect Flows*.

Em um segundo momento, visando buscar um ainda melhor desempenho do método, foram analisadas as implementações de extração das *features* de domínio da frequência presentes no livro *Introduction to Audio Analysis: A MATLAB Approach* [22]. Verificando que algumas delas não necessitavam do valor da frequência como entrada, decidiu-se implementar a extração delas, aplicando a TDF sobre o *Aspect Flow*, levando a um fictício domínio de frequência (já que, na verdade, um *Aspect Flow* não advém de uma onda).

Sendo assim, o método apresenta cinco *Audio-Like Features*, três adaptadas do domínio do tempo: *Energy*, *Median-Crossing Rate*, e *Energy Entropy*; e duas adaptadas do domínio da frequência: *Spectral Entropy e Spectral flux* (o termo *Spectral* foi mantido para mais fácil referência às *features* originais). Em seguida, as *features* serão detalhadas.

4.3.1 *Energy*

Como a versão original, essa *feature* reflete a magnitude total do aspecto no *Aspect Flow* ([27]).

Seja $x_i(n), n = 1, ..., F_L$ a sequência de sentenças do i-ésimo *frame*, onde F_L é o tamanho do *frame*. A implementação da Energy é definida pela Equação 4.1:

$$E(i) = \frac{1}{F_L} \sum_{n=1}^{F_L} |x_i(n)|^2 \tag{4.1}$$

Aqui a *Energy* foi normalizada pela divisão por F_L para remover a dependência do tamanho do *frame*. Quanto mais forte um aspecto aparece em um *frame*, maior o valor da *Energy* no mesmo.

4.3.2 *Median-Crossing Rate*

Median-Crossing Rate (MCR) é uma adaptação da *feature* de análise de áudio *Zero-Crossing Rate* (ZCR). Como no sinal de áudio a amplitude varia de -1 a 1, a ZCR é o número de vezes que o sinal muda de valor.

Como não é possível garantir que todas as entradas de representação de aspecto a serem tratadas pelo método sigam o intervalo da amplitude dos sinais de áudio $[-1, 1]$, esta implementação usa o *Aspect Flow Median* ($flowmedian$) como "linha" para calcular o número de vezes que será atravessada em um *frame*.

A MCR é definida de acordo com a Equação 4.2:

$$MCR(i) = \frac{1}{2F_L} \sum_{n=1}^{F_L} |msgn[x_i(n)] - msgn[x_i(n-1)]| \tag{4.2}$$

onde $msgn$ é a modificação da função sinal, a *Median Sign Function*, denotada pela Equação 4.3:

$$msgn[x_i(n)] = \begin{cases} 1, & \text{se } x_i(n) > flowmedian. \\ -1, & \text{se } x_i(n) < flowmedian. \\ 0, & \text{caso contrário.} \end{cases} \tag{4.3}$$

A MCR pode ser interpretada como uma medida de ruído do aspecto estudado em um *Aspect Flow*; em outras palavras, reflete o nível de variação do aspecto dentro do *frame*.

4.3.3 *Energy Entropy*

A *Energy Entropy* pode ser interpretada como uma medida de mudanças abruptas no nível de energia de um *Aspect Flow* (como na análise de áudio [22]). Por exemplo, ela pode detectar se um *frame* apresenta sentenças com diferenças profundas de níveis de subjetividade.

Para extraí-la, primeiro é necessário dividir cada *frame* em K *sub-frames*. Esse parâmetro não pode ser menor que $K = 2$ para garantir que vão existir pelo menos 2 *sub-frames*. Consequentemente, vai depender do tamanho médio de *frame* gerado em um *dataset*, assim, o método pode evitar o efeito indesejado de gerar muitos *sub-frames* em um *frame* curto.

Daí, para cada *sub-frame* j, é computada a *Energy* como na Equação (2.1) que, então é dividida pela *Energy* total do *frame*, $Eframe_i$. Essa divisão é necessária para tratar a sequência resultante de valores de *Energy* dos *sub-frames*, e_j, $j = 1, ..., K$, como uma sequência de probabilidades, como mostra a Equação 4.4:

$$e_j = \frac{Esubframe_j}{Eframe_i} \tag{4.4}$$

onde

$$Eframe_i = \sum_{k=1}^{K} Esubframe_k \tag{4.5}$$

Como passo final, a *Energy Entropy*, $Ent(i)$ da sequência e_j é computada de acordo com a Equação 4.6:

$$Ent(i) = -\sum_{j=1}^{K} e_j * log_2(e_j) \tag{4.6}$$

Quanto mais mudanças significativas um *frame* apresenta, menor o valor resultante da *Energy Entropy*.

4.3.4 Spectral Entropy

A *Spectral Entropy* é o cálculo da entropia da energia espectral do *frame* e é calculada de forma similar à *Energy Entropy*.

Primeiramente, divide-se novamente cada *frame* em K *sub-frames*, sendo necessário o mesmo cuidado com esse parâmetro já discutido quando foi apresentada a *Energy Entropy*.

Daí, a *Spectral Energy* $SEsubframe_f$ do f-ésimo *sub-frame*, $f = 1, ..., L-1$, é então normalizada pela *Spectral Energy* total do *frame*, dada pela Equação 4.7:

$$se_f = \frac{SEsubframe_f}{\sum_{f=0}^{L-1} SEsubframe_f}, f = 1, ..., L-1 \tag{4.7}$$

E, então, a *Spectral Energy* normalizada se_f é então computada pela Equação 4.8:

$$SEnt(i) = -\sum_{f=0}^{L-1} se_f * log_2(se_f) \tag{4.8}$$

A interpretação desta *feature* é semelhante à da *Energy Entropy*, apenas modificando o domínio.

4.3.5 Spectral Flux

A *feature Spectral Flux* mede as mudanças que ocorrem no espectro considerando dois *frames* sucessivos. Novamente, considere $X_i(k), k = 1, ..., Wf_L$ como sendo a magnitude dos coeficientes do i-ésimo *frame* resutante da TDF. A *Spectral Flux* é computada pelo quadrado da diferença entre as magnitudes de espectro normalizadas de dois *frames* consecutivos, como mostra a Equação 4.9:

$$SFlux_{(i,i-1)} = \sum_{k=1}^{Wf_L} \left(EN_i(k) - EN_{i-1}(k)\right)^2 \tag{4.9}$$

onde $EN_i(k)$ é o k-ésimo coeficiente da TDF normalizada do i-ésimo *frame*, calculado como mostra a Equação 4.10.

$$EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} (X_i(l)} \tag{4.10}$$

Um maior valor dessa *feature* leva a uma maior variação entre dois *frames* consecutivos, podendo capturar informações bem descritivas sobre o comportamento do aspecto no texto.

Como é possivel perceber na Equação (4.2), para que seja corretamente calculada a MCR, um *Aspect Flow* deve conter, pelo menos, duas sentenças por *frame*. Por sua vez, para ser possível extrair corretamente as *features Energy Entropy* e *Spectral Entropy* (Equações (4.6) e 4.8)), um *Aspect Flow* deve conter, pelo menos, uma sentença por *sub-frame*. Ainda assim, essas *features* se tornam mais descritivas quando há mais sentenças por *sub-frame*,

uma vez que geram mais valores para o cálculo da *Energy* e *Spectral Energy* do *sub-frame*. Considerando tudo isso, este método não é apropriado para analisar pequenos textos, como os advindos de microblogs.

4.4 Considerações Finais

Neste capítulo foi apresentado o método proposto para representação e análise de aspectos linguísticos em textos com foco na manutenção da forma como os aspectos são explorados ao longo do texto. O método é inspirado na área de análise de áudio, trazendo para o PLN conceitos como divisão em *frames* e extração de *features* típicas daquela área. Durante o capítulo foi explanada a forma de implementação do método, incluindo como são calculadas as *Audio-Like Features*. De acordo com as equações que regem a extração das *features*, é possível afirmar que o método não é adequado para análise de pequenos textos.

No próximo capítulo, serão descritos alguns dos experimentos realizados com objetivo de verificar a eficácia do método em tarefas de classificação, bem como, serão apresentados e discutidos os resultados desses experimentos.

Capítulo 5

Avaliação Experimental e Discussão

Este capítulo apresenta a avaliação experimental do método proposto, descrevendo a configuração dos diferentes experimentos realizados e discutindo os resultados obtidos, objetivando responder as questões de pesquisa levantadas no Capítulo 1 e abaixo revisitadas:

- QP1 - É possível criar um método de análise de textos inspirado na forma como a análise de áudio é realizada, amoldando a extração de *features* para o novo domínio?
- QP2 - É possível obter boa eficácia do método inspirado em análise de áudio em tarefas de classificação textual ao usar, como entrada, fluxos construídos a partir de distâncias semânticas entre texto e léxicos que remetem a um certo aspecto?
- QP3 - O uso de *frames* e *features* adaptadas do domínio de análise de áudio ocasiona boa eficácia em tarefas de classifição de textos?
- QP4 - Quando testada em várias tarefas de classificação de textos, a análise adaptada do domínio de áudio demonstra boa eficácia em todas delas?
- QP5 - O uso de *features* que refletem a variação observada nos fluxos promovem melhores resultados que *features* sintetizadoras em tarefas de classificação textual?
- QP6 - Caso a resposta à questão QP5 seja afirmativa, quão mais eficaz é utilizar *features* que refletem a variação observada nos fluxos quando comparado à utilização de *features* sintetizadoras em tarefas de classificação textual?

- QP7 - O uso de *features* extraídas dos fluxos promovem melhores resultados que o uso dos próprios fluxos em tarefas de classificação textual?

- QP8 - Caso a resposta à questão QP7 seja afirmativa, quão mais eficaz é utilizar *features* extraídas de fluxos quando comparado à utilização dos próprios fluxos em tarefas de classificação textual?

Os experimentos abordam três tarefas de classificação, a saber: classificação de notícias falsas (em inglês e português) baseada no aspecto subjetividade - essas tarefas tem por objetivo diferenciar notícias falsas de notícias legítimas; classificação de artigos de colunas de jornais em português, também baseada no aspecto subjetividade - esta tarefa, por sua vez, tem o intuito de distinguir artigos de colunas de jornais de notícias objetivas, que relatam fatos; e classificação de avaliações de filmes em inglês, baseada no aspecto sentimento - neste caso, o propósito é dicernir entre avaliações positivas e negativas. Os experimentos que abordam a classificação de notícias falsas serão abordados em seções diferentes para inglês e português, facilitando a organização e entendimento do texto, uma vez que envolvem *datasets* e léxicos diferentes.

Para efeitos de comparação, os experimentos também foram executados utilizando dois outros métodos: o primeiro é o método proposto por Jeronimo et al. [29], que utiliza *features* sintetizadoras para realizar classificação; o segundo é a utilização dos próprios *Aspect Flows* (dos quais as *Audio-Like Features* são extraídas) como *features* para classificação.

Os resultados apresentados neste capítulo se referem aos experimentos realizados com a versão mais atual do método, em que são extraídas as cinco *features* inspiradas na área de análise de áudio descritas no Capítulo 4. Entretanto, durante o processo de desenvolvimento deste trabalho, esses e outros experimentos foram realizados, utilizando apenas as *features* inspiradas no domínio do tempo. Alguns desses experimentos estão relatados no artigo publicado na Conferência LREC 2020 [56] (Apêndice A).

5.1 Classificação de Notícias Falsas em Inglês

Dada a disseminação massiva de notícias falsas propiciada pelas mídias sociais e aplicações de troca de mensagens e as grandes consequências a que essa disseminação pode levar, a

necessidade de se detectar notícias falsas é cada dia mais presente[1].

Habitualmente, documentos comprometidos em relatar fatos de forma verídica e imparcial, como notícias de jornais confiáveis, inclinam-se a usar uma linguagem mais objetiva, evitando palavras e expressões que denotem sentimento ou um tom mais argumentativo. Por outro lado, documentos que buscam persuadir o leitor, tendem a usar uma linguagem mais subjetiva [37; 59]. É neste caso que se encaixam as notícias falsas, que tentam convencer o leitor de algo que não é verídico. Porém, ainda assim, as notícias falsas almejam se passar por legítimas, logo, elas podem explorar o nível de subjetividade de forma variável ao longo do texto. Por exemplo, uma notícia falsa pode trazer apenas informações verídicas no início do texto, utilizando linguagem objetiva e, em seguida, incluir informações falsas através de linguagem mais subjetiva, mantendo o objetivo de passar credibilidade e, assim, se passar por notícias legítimas.

Diante do exposto, decidiu-se realizar experimentos utilizando o método proposto para avaliar seu desempenho na tarefa de classificação de notícias falsas, almejando diferenciá-las de notícias legítimas quanto ao aspecto subjetividade.

5.1.1 Descrição do Experimento

Dataset

O *dataset* utilizado foi compilado e disponibilizado por Jeronimo et al. [32] e engloba 5.994 notícias legítimas e 218 notícias falsas escritas em inglês. As notícias legítimas foram retiradas do *All The News Dataset* disponível no Kaggle[2], sendo 2.598 provenientes da CNN[3], 1.798 do The Guardian[4] e 1.598 do The New York Times[5], publicadas entre os anos de 2016 e 2017. As notícias falsas, por sua vez, foram compiladas por Torabi ASR e Taboada [7], sendo 103 notícias políticas provenientes do Snopes[6], 75 notícias de política provenientes do trabalho de Horne e Adali [25] e 40 notícias provenientes da notícias falsas mais bem

[1] https://www.acritica.com/channels/coronavirus/news/crenca-nas-fake-news-potencializa-disseminacao-e-problemas-da-pandemia

[2] https://www.kaggle.com/snapcrack/all-the-news

[3] www.cnn.com

[4] www.theguardian.com

[5] www.nytimes.com

[6] https://github.com/sfu-discourse-lab/

ranqueadas do Buzzfeed[7].

Léxicos de Subjetividade

Como fonte de informação sobre o aspecto a ser usado para criação dos fluxos, neste experimento foram usados, concomitantemente, três diferentes conjuntos de léxicos que objetivam refletir diferentes dimensões de subjetividade em inglês.

O primeiro conjunto foi compilado por Recasens et al. [49] e compreende seis diferentes dimensões de termos que tendem a induzir viés em textos, denomidados pelos autores de "*bias-inducing*", a saber:

- *Factive Verbs*: pressupõem a verdade de uma cláusula de complemento. Exemplos dos 27 termos que a compõem: *realize, forget, exciting*.
- *Implicative Verbs*: implica a verdade ou falsidade da cláusula de compemento. Exemplos dos 32 termos que a compõem: *succeed, fail, neglect*.
- *Assertive verbs*: verbos que seus complementos afirmam uma proposição. Exemplos dos 66 termos que a compõem: *believe, figure, affirm*.
- *Hedges*: usados para reduzir o comprometimento com a verdade de uma proposição. Exemplos dos 100 termos que a compõem: *apparently, could, estimate*.
- *Reporting Verbs*: geralemnte usados para descrever atividades ou ações de uma terceira pessoa. Exemplos dos 181 termos que a compõem: *accuse, assure, claim*.
- *Bias-inducing lemmas*: exemplos dos 654 termos que a compõem: *advocate, amazing, barbarian*.

O segundo conjunto de léxicos foi apresentado por Wilson et al. [60] e é parte do projeto Multi-Perpective Question Answering (MPQA) Subjectivity Lexicons[8]. Esse conjunto é dividido em polaridades de sentimentos (positiva e negativa - duas dimensões) e classificado em subjetividade forte e fraca. Foram utilizados apenas os termos pertencentes à categoria

[7]https://github.com/BuzzFeedNews/ 2017-12-fake-news-top-50

[8]https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

de subjetividade forte de ambas as polaridades, resultando em 3.078 termos de polaridade negativa e 1.482 de polaridade positiva.

O terceiro conjunto foi proposto por Deng et al. [13], baseado em um tipo de inferência de opinião que surge quando são expressadas opiniões em relação a eventos, gerando efeitos positivos ou negativos relativos aos eventos. Os termos, extraídos de documentos que apresentam subjetividade como blogs e editoriais, apresentam polaridade (negativa ou positiva) em duas categorias: Gold Standard, extraídos por anotadores; e EffectWordNet, extraídos automaticamente pelo método proposto por Deng et al.[13]. Foram utilizados apenas os termos da categoria Gold Standard, que apresenta 1.003 termos de polaridade negativa e 493 de polaridade positiva.

Experimento

O principal objetivo do experimento é avaliar a efetividade da análise baseada em *Audio-Like Features* a partir da representação das notícias por fluxos de subjetividade (*Aspect Flows*) para a classificação de notícias falsas.

Para a construção dos fluxos de subjetividade, inicialmente foi realizada uma etapa de pré-processamento em que, com o texto já fragmentado em sentenças, foi realizada a remoção das *stop words* e de sentenças que continham até duas palavras. Posteriormente, foram calculadas as distâncias WMD de cada sentença das notícias para as dez dimensões dos três conjuntos de léxicos de subjetividade descritas anteriormente, no espaço de *embedding*. Logo, para cada notícia do *dataset* são criados dez fluxos de subjetividade, um para cada dimensão. Daí, cada fluxo é fragmentado em *frames* e são calculadas as *Audio-Like Features* para cada *frame* do fluxo. O modelo de *word embedding* utilizado foi o amplamente difundido conjunto de *word vectors* pré-treinados do Word2Vec[9].

Configuração do Experimento

Primeiramente, foi avaliada a média do número de sentenças por documento do *dataset* para que fosse definido o número de *frames* em que os fluxos de subjetividade (*Aspect Flows* deste experimento) seriam divididos e, também, o valor que deveria ser escolhido para o parâmetro K, necessário para o cálculo da *Energy Entropy e da Spectral Entropy*. As notícias legítimas

[9]https://code.google.com/archive/p/word2vec/

apresentam uma média de, aproximadamente, 41 sentenças por documento, enquanto as falsas apresentam uma média de, aproximadamente, 22 sentenças por documento. Diante disso, decidiu-se que os fluxos seriam divididos em **3 *frames***, resultando em 13,66 e 7,3 sentenças por *frame*, na média, para as notícias legítimas e falsas, respectivamente. O valor do parâmetro K foi definido como $K = 2$, pois, dessa forma, haveria, pelo menos três sentenças por *sub-frame*, em média. Convém relembrar que, para o correto cálculo da *Energy Entropy* e da *Spectral Entropy*, o número mínimo é de uma sentença por *sub-frame*, mas para *features* mais significativas, ter mais sentenças por *sub-frame* é essencial.

Considerando essas decisões, ainda assim, alguns documentos não cumpriram o requisito mínimo de uma sentença por *sub-frame*. Nestes casos, foi aplicada uma técnica de preenchimento que consiste em repetir o valor obtido na última sentença de cada *frame* até que este alcance o tamanho mínimo de $2K$ sentenças. Assim, é possível ter, pelo menos, duas sentenças por *sub-frame*. Dessa forma, torna-se possível não apenas o correto cálculo das *features*, mas também o cálculo de *features* mais significativas (com mais de uma sentença por sub-fame), ainda mantendo o cuidado de evitar muitas modificações no *Aspect Flow*. Essa técnica foi denominada de *Last Frame Sentence Padding*.

A Figura 5.1 ilustra a aplicação da *Last Frame Sentence Padding* a um *Aspect Flow* de 7 sentenças que precisa ser dividido em 3 *frames* com $K = 2$. Assim, depois de dividido em *frames*, o *Aspect Flow* apresenta 3, 2 e 2 sentenças no *Frame* 0, *Frame* 1 e *Frame* 2, respectivamente. Logo, como $K = 2$, cada *frame* deve apresentar 4 sentenças para o cálculo de todas as *Audio-Like Features*, sendo o valor da última sentença do *Frame* 0 repetido uma vez, e último valor dos demais *frames*, duas vezes. Em casos mais raros em que o texto não apresente no mínimo o número de sentenças igual ao de *frames*, o valor da última sentença do texto é repetido até o final. Essa técnica permite não apenas que sejam extraídas todas as *Audio-Like Features* corretamente, mas também que sejam mantidas características essenciais do texto que o método almeja analisar. Por exemplo, ao realizar o preenchimento por *frame*, é possível preservar o posicionamento de cada porção do texto: sentenças do meio e do fim do texto se mantêm nessa posição, o que nem sempre ocorreria se o preenchimento fosse realizado apenas no fim do texto. Repetir o valor da última frase também permite manter características que o autor quis imprimir naquela porção do texto, como uma continuidade do conteúdo apresentado, o que não ocorreria se, por exemplo, fosse realizado o

preenchimento com um valor pré-determinado.

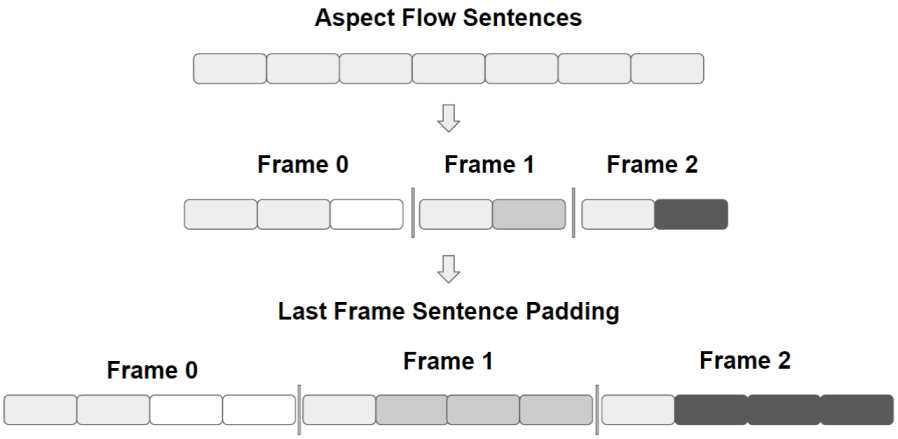

Figura 5.1: *Last Frame Sentence Padding*.

Como classificadores para avaliação da eficácia do método proposto, foram utilizados os modelos *Random Forest (RF)* e *XGBoost (XGB)*. Para fins de comparação, foi realizada a classificação alimentando os modelos supracitados com quatro grupos distintos de *features*: (1) as *features* sintetizadoras propostas por Jerônimo et al. [29], aqui denominadas de "*Average Model*" (AVGM); (2) os valores de WMD obtidos para cada texto em relação a cada léxico, os *Aspect Flows*, como *features*, o qual será denominado de *WMD Flow Model* (WMDM); (3) as *Audio-Like Features* de todos os *frames* - "*All Frames Model*" (ALFM) referentes ao método proposto; e, finalmente, (4) as *Audio-Like Features* por *frame* - "*Single Frame Model*" com o número referente ao *frame* (SFM#), usadas para analisar o desempenho do método proposto considerando cada *frame* separadamente.

Para executar a classificação utilizando o WMDM, foi necessário realizar um preenchimento nos fluxos, uma vez que os textos possuem tamanhos diferentes. Neste caso, foram dispensados os textos não pertencentes ao 0,9-percentil, pois tinham tamanho bem diferente dos demais, o que iria causar um preenchimento muito volumoso nos outros textos. Em seguida, foi realizado o prenchimento utilizando o último valor de cada fluxo até o tamanho do maior texto dentro do percentil, que possui 67 sentenças.

Dado o desbalanceamento entre o número de textos de cada classe e o maior interesse com a classe das notícias falsas, minoritária, esta foi configurada como a classe positiva.

Pelo mesmo motivo, para avaliar os modelos, foram utilizadas, além das métricas *Precision* e *Recall*, métricas que, de alguma forma, refletem uma combinação entre elas. A primeira é o *F1-Score*, que combina *Precision* e *Recall* através da média harmônica entre elas. A segunda é a *Area Under the Precision-Recall curve (PR-AUC)*, particularmente apropriada para análise de cenários em que há grande desbalanceamento de classes, como discutido nos trabalhos [50; 14], pois calcula a área abaixo da curva PR, esta gerada a partir dos valores de *Precision* a cada limiar de *Recall*, propiciando uma análise mais global do desempenho do classificador na tarefa.

É notório que, em *datasets* que tratam de notícias falsas, o número de notícias legítimas presentes é significativamente maior que o de notícias falsas, gerando grande desbalanceamento. Em cenários como esse, é comum o uso de técnicas de *oversampling*. Entretanto, assim como realizado porJerônimo et al. [29], neste experimento, foi utilizada a proporção de quatro notícias legítimas para uma notícia falsa, baseada no estudo de Silverman [52], que indica a presença dessa proporção na disseminação de notícias durante as eleições dos Estados Unidos de 2016. Além disso, o experimento foi executado 500 vezes, com randomização dos conjuntos de treino e teste das notícias legítimas e falsas, para que não se restringisse o uso a apenas uma pequena porção das diversas notícias legítimas disponíveis e fossem variados os textos de ambos os tipos de notícia como treino e teste a cada rodada.

5.1.2 Resultados e Discussão

A Tabela 5.1 apresenta as médias e desvios-padrão dos resultados das métricas *Precision, Recall, F1-Score* e *PR-AUC* das 500 rodadas de cada um dos modelos treinados.

Através desses resultados, é possível visualizar que todos os modelos que envolvem as *Audio-Like Features* apresentam melhores resultados que o AVGM e o WMDM em todas as métricas, tanto no *Random Forest* quanto no *XGBoost*. Em destaque na tabela, o valor de *Precision* do *XGBoost* apresentado pelo ALFM, de valor $0,79$, sendo $51,9\%$ maior que o AVGM e $25,39\%$ maior que o WMDM. Ainda no tocante à *Precision* do *XGBoost*, os modelos SFM0 e SFM2 também apresentam resultados expressivos, próximos ao do ALFM e maiores que o AVGM e WMDM em $46,15\%$ e $20,63\%$, respectivamente. Uma outra métrica destacada é a *PR-AUC* do *XGBoost* também atingida pelo ALFM, que apresenta valor de $0,53$, ou seja, $65,62\%$ e $51,43\%$ maior que o AVGM e o WMDM, respectivamente.

Tabela 5.1: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação de notícias falsas em inglês.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,50 \pm 0,14$	$0,53 \pm 0,16$	$0,76 \pm 0,10$	$0,73 \pm 0,10$	$0,73 \pm 0,11$	$0,71 \pm 0,10$
	XGB	$0,52 \pm 0,13$	$0,63 \pm 0,18$	$\mathbf{0,79 \pm 0,08}$	$0,76 \pm 0,10$	$0,74 \pm 0,10$	$0,76 \pm 0,09$
Recall	RF	$0,11 \pm 0,04$	$0,08 \pm 0,03$	$0,20 \pm 0,04$	$0,20 \pm 0,04$	$0,18 \pm 0,04$	$0,18 \pm 0,04$
	XGB	$0,10 \pm 0,03$	$0,07 \pm 0,03$	$0,21 \pm 0,03$	$\mathbf{0,22 \pm 0,04}$	$0,18 \pm 0,03$	$0,21 \pm 0,03$
F1-Score	RF	$0,18 \pm 0,05$	$0,13 \pm 0,05$	$0,31 \pm 0,06$	$0,31 \pm 0,05$	$0,29 \pm 0,06$	$0,28 \pm 0,06$
	XGB	$0,17 \pm 0,05$	$0,13 \pm 0,05$	$0,34 \pm 0,05$	$0,34 \pm 0,05$	$0,29 \pm 0,05$	$0,33 \pm 0,04$
PR-AUC	RF	$0,31 \pm 0,03$	$0,28 \pm 0,03$	$0,49 \pm 0,04$	$0,46 \pm 0,04$	$0,46 \pm 0,04$	$0,47 \pm 0,04$
	XGB	$0,32 \pm 0,04$	$0,35 \pm 0,04$	$\mathbf{0,53 \pm 0,03}$	$0,49 \pm 0,04$	$0,49 \pm 0,04$	$\mathbf{0,53 \pm 0,04}$

Ainda referente a esta métrica, há o fato de que o SFM2 apresenta o mesmo resultado que o ALFM. Um outro destaque que concerne a modelos referentes a *frames* individuais é o valor de *Recall* do XBoost apresentado pelo SFM0, $0,22$, sutilmente maior que o apresentado pelo ALFM, contudo mais que o dobro e o triplo do valor apresentado pelos AVGM e WMDM, respectivamente.

Esses resultados mostram que, para esta tarefa, é possível analisar textos usando *features* inspiradas na análise de áudio extraídas de fluxos construídos a partir de distâncias semânticas entre texto e léxico e, ainda, obter bons resultados, respondendo à questão de pesquisa QP1 e, parcialmente, à QP2 e QP3. Pela comparação dos resultados dos modelos treinados com *Audio-Like Features* e o AVGM, percebe-se que o uso de *features* extraídas de fluxos apresenta, neste caso, maior eficácia que *features* sintetizadoras com as margens recentemente citadas, respondendo positivamente às questões QP5 e QP6. O mesmo acontece ao serem comparados os resultados dos modelos quando treinados com as *Audio-Like Features* e o WMDM, respondendo positivamente à QP7 e as margens discutidas respondem à QP8.

Um outro fato possível de se perceber nesses resultados, considerando apenas os modelos envolvendo *Audio-Like Features*, é que, em geral, o SFM0 e SFM2 apresentam um melhor resultado, mesmo que sutil, que o SFM1 na classificação. Ressaltando, novamente, que em alguns resultados obtiveram patamar igual ou levemente superior que o ALFM. Esses fatos podem sugerir que o aspecto explorado, a subjetividade, é mais decisivo ao diferenciar notícias legítimas e falsas deste *dataset* nas porções inicial e final do texto. Assim sendo,

é possível perceber que o método proposto pode ser capaz de apontar trechos dos textos que apresentam informações mais significativas referentes à tarefa em questão. Ressalta-se que, neste caso, a análise de apenas um desses trechos supera a eficácia dos modelos de comparação (AVGM e WMDM).

Análise de *Audio-Like Features*

Para exemplificar o tipo de informações sobre como o aspecto subjetividade é explorado no *dataset* o método proposto pode oferecer, será apresentada a análise de algumas *Audio-Like Features* ao longo dos três *frames*.

A Figura 5.2 apresenta os boxplots da Spectral Entropy do fluxo "Bias-inducing Lemma" dos *frames* 0, 1 e 2. Como é possível nela observar, as notícias falsas apresentam menores valores de *Spectral Entropy* que as legítimas em todos os *frames*, ou seja, as notícias falsas sofrem mais mudanças abruptas nas distâncias WMD entre sentenças e a dimensão *bias* avaliadas sob a perspectiva do domínio da frequência, apresentando mais variação no nível de subjetividade referente à referida dimensão. Também é possível observar que os valores das notícias falsas vão crescendo no decorrer dos *frames*, ou seja, ao longo do texto. Enquanto as notícias legítimas apresentam uma diminuinção no *frame* 1, aproximando-se mais das notícias falsas (bem perceptível na observação das medianas), mostrando que essa *feature* demonstra menor poder de diferenciação entre as classes de notícia na porção medial do texto.

Por sua vez, a Figura 5.3 apresenta os boxplots da *feature Spectral Flux* do fluxo "Hedges". Como essa *feature* mede as mudanças ocorridas considerando dois *frames* consecutivos, o *frame* 0 sempre apresenta este comportamento, pois recebe o valor 0 para esta *feature*, uma vez que ainda não há outro *frame* para o cálculo. Percebe-se que as notícias falsas sofrem menor variação do comportamento da referida dimensão entre seus *frames*, visto que apresentam menor valor que as notícias legítimas em ambos os *frames*. Entretanto, é importante frisar que enquanto o valor da *Spectral Flux* das notícias falsas diminui consideravelmente do *frame 1* para o *frame* 2, o valor das legítimas tem um leve crescimento, levando a uma maior discrepância dessa *feature* entre as classes de notícia no *frame* 2.

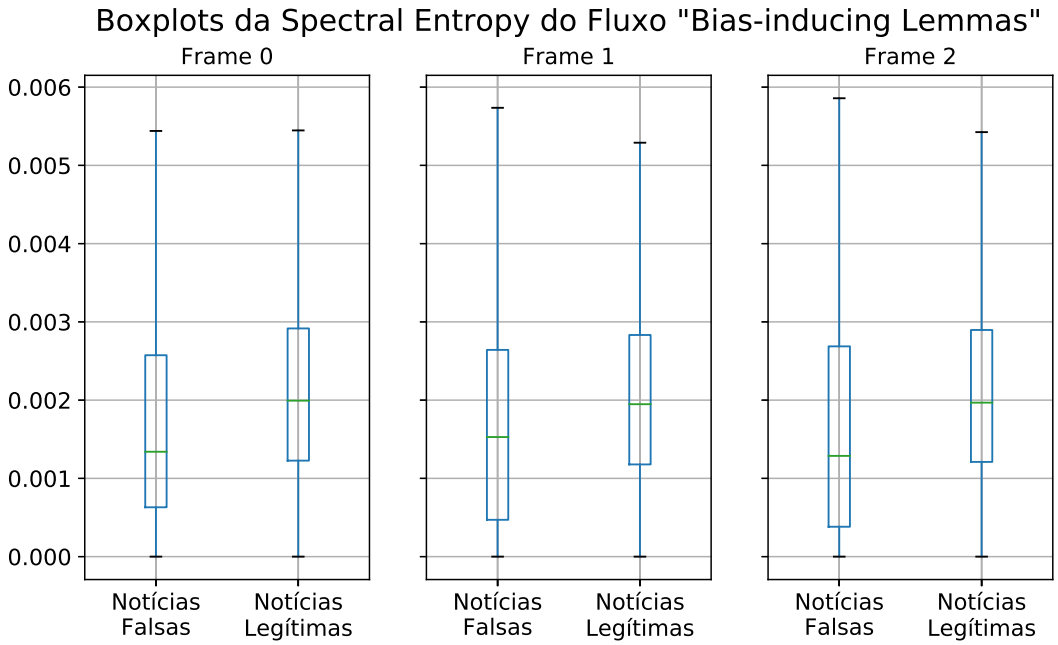

Figura 5.2: Boxplots da *feature Spectral Entropy* do Fluxo "*Bias-inducing Lemmas*".

5.2 Classificação de Notícias Falsas em Português

Foram conduzidos também experimentos envolvendo classificação de notícias falsas, mas agora tratando com textos em português.

5.2.1 Descrição do Experimento

Dataset

O *dataset* utilizado engloba notícias legítimas e falsas brasileiras. É o mesmo utilizado no trabalho de Jeronimo et al. [29], mas com a inclusão de 26 notícias falsas ao conjunto inicial.

O *dataset* apresenta um total de 207.914 notícias legítimas publicadas entre os anos de 2014 e 2017, coletadas de dois dos maiores *sites* de notícias brasileiros, Estadão[10] e Folha de São Paulo[11] e divididas em quatro diferentes domínios: Cultura, Economia, Esportes e Política.

O *dataset* é composto, ainda, por 121 notícias falsas, originadas de várias fontes, que

[10]https://www.estadao.com.br/
[11]https://www.folha.uol.com.br/

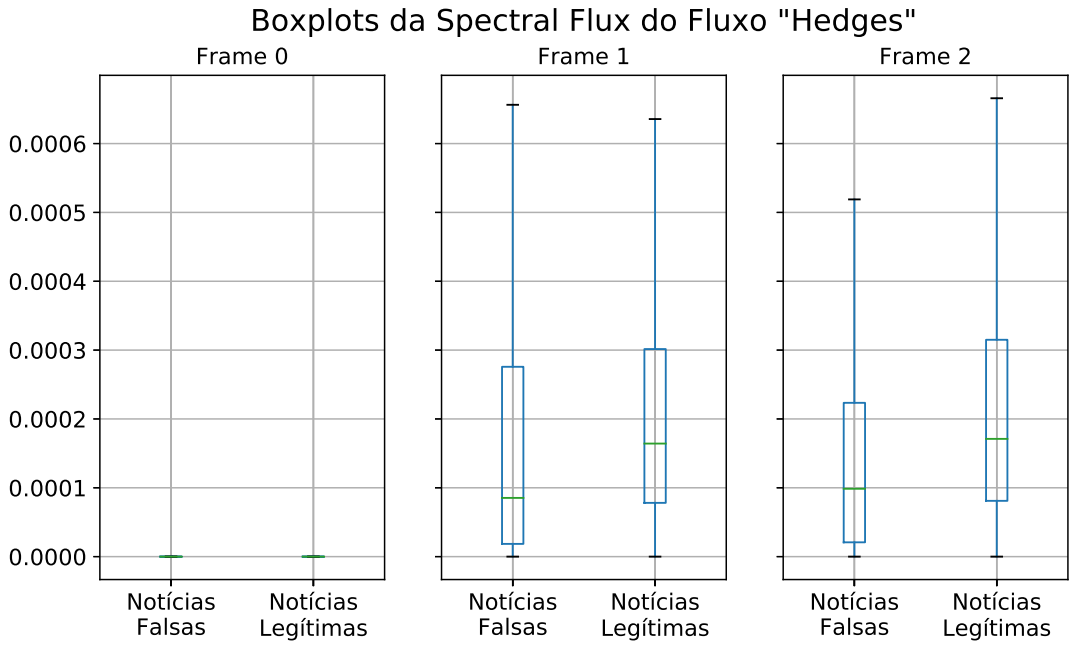

Figura 5.3: Boxplots da *feature Spectral Flux* do Fluxo "*Hedges*".

foram muito disseminadas no Brasil entre 2010 e 2017, coletadas em dois populares *sites* de checagem de fatos, o e-Farsas[12] e o Boatos[13].

Léxico de Subjetividade

Como fonte de informação sobre o aspecto a ser usado para criação dos fluxos, foi empregado o léxico de subjetividade em português do Brasil apresentado no trabalho de Amorim et al. [4]. Esse léxico foi criado por linguistas brasileiros e é dividido em cinco dimensões, descritas a seguir:

- A dimensão argumentação representa palavras e expressões relacionadas a um discurso mais argumentativo. Esse tipo de discurso é frequentemente utilizado quando alguém tenta convencer outra pessoa de um ponto de vista específico. Exemplos das 116 palavras e expressões que a compõem: "ao menos", "o único", "mas também" e "por essa razão".

- A dimensão pré-suposição inclui termos relacionados a uma suposição prévia sobre

[12] http://www.e-farsas.com/
[13] http://www.boatos.org/

algo. É um tipo de discurso utilizado principalmente em situações em que o interlocutor assume algo como verdade, mesmo que este não seja o caso. Exemplos das 54 palavras e expressões que a compõem: “para demonstrar”, “descobrir”, “adivinhar”, “entende”.

- A dimensão sentimento contém palavras e termos relacionados a um discurso emocional, quando o autor tenta envolver emocionalmente o leitor. Exemplos das 151 palavras e expressões que a compõem: “amar”, “constranger”, “desapontar”, “encorajar”.

- A dimensão de valoração expressa palavras relacionadas a intensificação de algo. Exemplos das 81 palavras e expressões que a compõem: “escasso”, “excessivo”, “incrível”, “lamentável”.

- A dimensão modalização reúne palavras e termos utilizados quando o interlocutor tem um posicionamento estabelecido sobre algo ou alguém. Exemplos das 55 palavras e expressões que a compõem: “aconselhar”, “indubitável”, “presumir”, “supor”.

Experimento

Assim como no experimento com notícias em inglês, este experimento objetiva avaliar a efetividade da representação das notícias por fluxos de subjetividade e da análise baseada em *Audio-Like Features* para a classificação de notícias falsas.

Para a construção dos fluxos de subjetividade, também foi seguida a mesma metodologia que o experimento com notícias em inglês: após o pré-processamento (remoção de *stop words* e sentenças com até duas palavras), foram calculadas as distâncias WMD de cada sentença das notícias para as cinco dimensões do léxico de subjetividade no espaço de *embedding*. Logo, para cada notícia do *dataset* são criados cinco fluxos de subjetividade, um para cada dimensão. Daí, cada fluxo é fragmentado em *frames* e são calculadas as *Audio-Like Features* para cada *frame* do fluxo. O modelo de *word embedding* utilizado foi criado a partir de um grande volume de artigos da Wikipedia em português disponibilizado por Jerônimo et al. [29].

Inspirado no artigo de Jeronimo et al. [29], foram executados experimentos em vários cenários, descritos a seguir:

1. **Notícias Legítimas x Notícias Falsas**: Neste cenário, o conjunto de dados de treino e de teste são constituídos por notícias legítimas de quaisquer domínios e fontes a que pertencem e todas as notícias falsas, que também são provenientes de uma variedade de fontes e domínios.

2. ***Cross-domain***: Neste cenário, são escolhidos domínios diferentes das notícias legítimas para formar os conjuntos de treino e teste no que diz respeito a essa classe de notícias. Por exemplo, o classificador pode ser treinado com notícias legítimas de Cultura e testado com notícias legítimas de Economia.

3. ***Cross-source***: Neste cenário, o conjunto de treino é formado por notícias do Estadão e o de teste, por notícias da Folha de São Paulo, no tocante a notícias legítimas.

4. ***Cross-source-domain***: Este cenário é uma mescla dos dois anteriores, ou seja, apenas notícias do Estadão são utilizadas para treino e da Folha de São Paulo para teste, porém são variados também os domínios de treino e teste. Por exemplo, o classificador pode ser treinado com notícias legítimas de Cultura do Estadão e testado com notícias legítimas de Economia da Folha de São Paulo.

Configuração do Experimento

Seguindo o mesmo procedimento do experimento com notícias em inglês, foi avaliada a média do número de sentenças por documento do *dataset* para que fosse definido o número de *frames* em que os *Aspect Flows* seriam divididos e o valor do parâmetro K. As notícias legítimas apresentam uma média de 21 sentenças por documento, enquanto as falsas apresentam uma média de 14 sentenças por documento. Diante disso, decidiu-se que os fluxos seriam novamente divididos em **3 *frames***, resultando em 7 e 4,67 sentenças por *frame*, na média, para as notícias legítimas e falsas, respectivamente. Desde que foram obtidos frames com média de poucas sentenças, o valor do parâmetro K foi definido como $K = 2$, para que houvesse, na média, pelo menos duas sentenças por *sub-frame*, próximo ao número mínimo necessário para o cálculo da *Energy Entropy e da Spectral Entropy* corretamente. Daí, foi aplicada a técnica *Last Frame sentence Padding* aos documentos que não cumpriram o requisito de duas sentenças por *sub-frame*.

Também foram mantidas outras configurações do experimento com notícias em inglês: os modelos de classificação, *Random Forest* e *XGBoost*; os grupos de *features* para avaliação e comparação, AVGM, WMDM, ALFM, SFM0, SFM1 e SFM2; as métricas de avaliação, *Precison, Recall, F1-Score* e *PR-AUC*. As notícias falsas são consideradas a classe positiva e a proporção de 4 notícias legítimas para uma notícia falsa é utilizada. Finalmente, 500 rodadas de classificação são executadas, havendo randomização dos textos que compõem o conjunto de treino e teste a cada rodada. Para executar a classificação utilizando o WMDM foi aplicado o mesmo preenchimento descrito na Seção 5.1.1, sendo que, neste caso, o tamanho do maior texto pertencente ao 0,9-percentil é 32 sentenças.

5.2.2 Resultados e Discussão

Nesta seção serão apresentados apenas os resultados do cenário em que não há distinção entre fontes e/ou domínios das notícias legítimas por dois motivos: (1) segundo Jeronimo et al. [29] é o cenário mais desafiador e (2) os resultados dos demais cenários não diferem muito dos que serão apresentados, não demonstrando uma condição muito melhor ou pior para que sejam discutidos. Os resultados dos demais cenários podem ser encontrados no Apêndice B.

A Tabela 5.2 apresenta as médias e desvios-padrão dos resultados das métricas *Precision, Recall, F1-Score* e *PR-AUC* das 500 rodadas de cada um dos modelos treinados.

Tabela 5.2: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação de notícias falsas em português.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,29 \pm 0,13$	$0,31 \pm 0,17$	$0,40 \pm 0,15$	$0,34 \pm 0,14$	$0,43 \pm 0,13$	$0,44 \pm 0,13$
	XGB	$0,31 \pm 0,13$	$0,38 \pm 0,16$	$\mathbf{0,46 \pm 0,14}$	$0,38 \pm 0,12$	$0,47 \pm 0,11$	$\mathbf{0,49 \pm 0,13}$
Recall	RF	$0,08 \pm 0,04$	$0,06 \pm 0,03$	$0,11 \pm 0,05$	$0,10 \pm 0,05$	$0,14 \pm 0,05$	$0,14 \pm 0,06$
	XGB	$0,08 \pm 0,03$	$0,08 \pm 0,04$	$0,15 \pm 0,05$	$0,13 \pm 0,05$	$\mathbf{0,18 \pm 0,05}$	$0,17 \pm 0,06$
F1-Score	RF	$0,12 \pm 0,05$	$0,10 \pm 0,05$	$0,17 \pm 0,07$	$0,15 \pm 0,07$	$0,21 \pm 0,07$	$0,21 \pm 0,07$
	XGB	$0,12 \pm 0,05$	$0,12 \pm 0,05$	$0,23 \pm 0,07$	$0,19 \pm 0,06$	$0,26 \pm 0,07$	$0,25 \pm 0,07$
PR-AUC	RF	$0,26 \pm 0,03$	$0,24 \pm 0,03$	$0,33 \pm 0,04$	$0,29 \pm 0,04$	$0,33 \pm 0,04$	$0,33 \pm 0,05$
	XGB	$0,27 \pm 0,03$	$0,31 \pm 0,04$	$\mathbf{0,39 \pm 0,05}$	$0,33 \pm 0,04$	$0,39 \pm 0,05$	$\mathbf{0,40 \pm 0,05}$

Os resultados apresentados mostram que o desempenho de todos os modelos neste *dataset* é bem aquém daquele obtido no experimento que envolve notícias em inglês. Isso pode ser devido à composição do *dataset* que contém um número bem maior de notícias legítimas e, principalmente, bem menor de notícias falsas, número este que delimita o tamanho dos conjuntos de treino e teste. Além disso, várias notícias falsas, são bem pequenas, implicando em forte aplicação da técnica de preenchimento (tanto para a extração das *Audio-Like Features*, quanto para o WMDM), o que pode influenciar no resultado da classificação.

Ainda assim, é possível visualizar que todos os modelos que envolvem as *Audio-Like Features* apresentam, novamente, melhores resultados que o AVGM e o WMDM em todas as métricas, tanto no *Random Forest* quanto no *XGBoost*. Em destaque na tabela, o valor de *Precision* do *XGBoost* apresentado pelo ALFM, de valor $0,46$, sendo $48,38\%$ maior que o AVGM e $21,05\%$ maior que o WMDM. Ainda relativo à *Precision* do *XGBoost*, é destacado o resultado do SFM2, $0,49$, que supera o resultado do ALFM e é maior que o AVGM e WMDM em $58,06\%$ e $28,95\%$, respectivamente. Uma outra métrica destacada é a *PR-AUC* do *XGBoost*, cujo resultado atingido pelo ALFM, de valor $0,39$, é maior que o AVGM e o WMDM em $44,44\%$ e $25,8\%$, respectivamente. Ainda referente a esta métrica, há o fato de que o SFM1 apresenta o mesmo resultado que o ALFM e o SFM2, um resultado levemente maior, $0,40$. Continuando com os valores destacados na tabela, percebe-se que o valor de *Recall* do XBoost apresentado pelo SFM1, $0,18$, 20% maior que o apresentado pelo ALFM e 125% maior que os valores apresentados pelo AVGM e WMDM.

Esses resultados mostram que, para classificação de notícias falsas em português, o uso das *Audio-Like Features* extraídas de fluxos construídos a partir de distâncias semânticas entre texto e léxico não é tão eficaz, respondendo, neste caso, negativamente às questões de pesquisa QP1, QP2 e QP3. Contudo, convém ressaltar a difícil natureza do *dataset* trabalhado. No entanto, é possível responder afirmativamente às questões de pesquisa QP5 e QP7, uma vez que o uso de *Audio-Like Features* apresenta maior eficácia que o AVGM e WMDM. As margens discutidas respondem às questões QP6 e QP8.

Ao analisar apenas os modelos envolvendo *Audio-Like Features*, em geral, o SFM2 apresenta melhor resultado na classificação que o SFM1 e, principalmente, que o SFM0. Em praticamente todos os resultados, o SFM2 obtive patamar igual ou superior que o ALFM. Esses fatos poderiam sugerir que o aspecto explorado, a subjetividade, é mais decisivo ao

diferenciar notícias legítimas e falsas deste *dataset* na porção final do texto. Entretanto, dada a discutida difícil natureza dos dados disponíveis, essa maior diferença no último *frame* também pode ter se dado pelo uso mais intenso de preenchimento das notícias falsas, que, mesmo que seja repetido o último valor, ao invés de inserido um valor pre-determinado, pode ajudar a diferenciar as notícias legítimas de falsas nesse trecho do texto.

Análise de *Audio-Like Features*

A Figura 5.4 apresenta os boxplots da *Energy* do fluxo "Argumentação"dos *frames* 0, 1 e 2. Como é possível nela observar, ambas as classes de notícias apresentam um crescimento da *Energy* no decorrer dos *frames*, ou seja, ao longo do texto. Logo, a presença do aspecto subjetividade relacionado ao léxico "Argumentação" é mais forte no decorrer do texto. É possível notar também que nos dois primeiros *frames* as notícias legítimas apresentam valores levemente maiores de *Energy*. Entretanto, essa situação se reverte no último *frame*, dado o crescimento acentuado dos valores das notícias falsas. Esse comportamento observado no último *frame* pode se dever à mais intensa aplicação do *Last Frame Sentence Padding* às notícias falsas deste *dataset*.

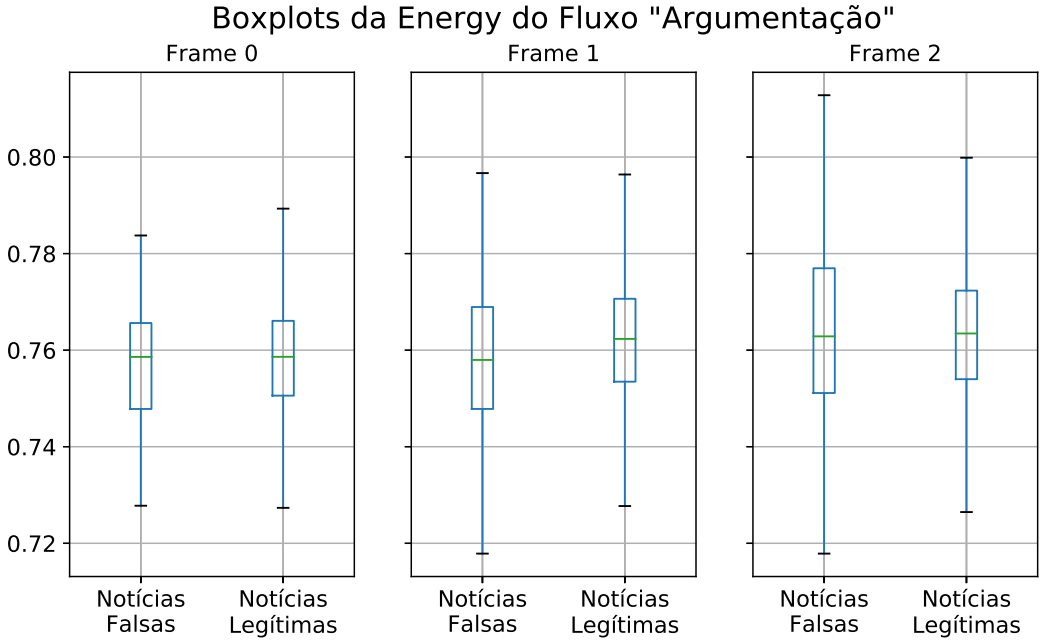

Figura 5.4: Boxplots da *feature Energy* do Fluxo "Argumentação".

Por sua vez, a Figura 5.5 apresenta os boxplots da MCR do fluxo de pré-suposição dos *frames* 0, 1 e 2. Pode-se perceber que as notícias falsas apresentam menor valor que as legítimas em todos os três *frames*. Além disso, os valores de MCR das notícias falsas apresentam um declínio no *frame* 1, voltando praticamente ao antigo patamar no *frame* 2, enquanto o declínio ocorre, no último *frame* em menor intensidade, quando consideradas as notícias legítimas. Isso mostra que as notícias legítimas são mais instáveis no tocante às distâncias à dimensão de pressuposição, mas as notícias falsas apresentam maior variação do valor de MCR entre os *frames*.

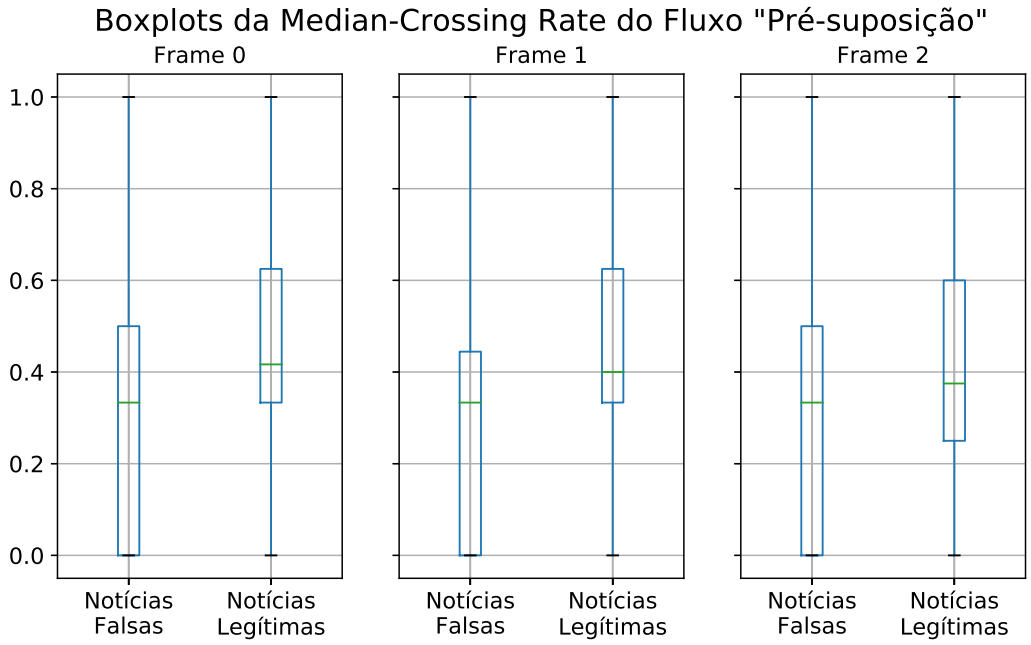

Figura 5.5: Boxplots da *feature Median-Crossing Rate* do Fluxo "Pré-suposição".

5.3 Classificação de Artigos de Colunas de Jornais

Colunas de jornais são seções dos jornais, geralmente temáticas, periódicas e escritas por um mesmo autor. Os artigos das colunas são normalmente caracterizados pela voz, personalidade e opiniões do autor, em oposição às notícias objetivas, por assim dizer, que reportam fatos.

Considerando essas características, parece viável diferenciar artigos de colunas de notícias baseado na subjetividade da linguagem usada nos textos. Logo, decidiu-se avaliar a

eficácia do método proposto ao realizar a tarefa de classificação entre notícias e artigos de colunas, usando uma metodologia bem similar à apresentada na Seção 5.1.

5.3.1 Descrição do Experimento

Dataset

Para este experimento, foram utilizadas as mesmas notícias legítimas do experimento de notícias falsas em português para representar as notícias, visto que não foram coletadas de colunas de jornais.

Para compor a representação dos artigos de colunas de jornais, foram coletados 7.062 artigos da Folha de São Paulo[14]. Os artigos coletados são provenientes de uma variedade de domínios (e.g., política, economia, negócio, turismo) e publicados entre os anos de 2010 e 2018.

Léxico de Subjetividade

Como as notícias e artigos utilizados estão escritos em português e o aspecto avaliado será a subjetividade, este experimento também utiliza o léxico de subjetividade proposto por Amorim et al. [4] para a criação dos fluxos.

Experimento

O principal objetivo do experimento é avaliar a eficácia da representação das notícias e artigos por fluxos de subjetividade e da análise baseada em *Audio-Like Features* para a classificação entre esses dois tipos de texto presentes em jornais.

A construção dos fluxos de subjetividade e extração das *Audio-Like Features* seguem os mesmos passos dos experimentos de classificação de notícias falsas: após a remoção de *stop words* e sentenças com até duas palvras, calculam-se as distâncias WMD de cada sentença dos textos para cada dimensão do léxico no espaço de *embedding*, usando o mesmo modelo de *word embedding* [29], gerando 5 fluxos para cada texto; estes, então, são fragmentados em *frames* e, daí, são extraídas as *Audio-Like Features* de cada *frame*.

[14]https://www.folha.uol.com.br/

Neste experimento, foi proposto apenas o cenário em que são utilizadas todas as notícias e artigos, independentemente de domínio e fonte.

Configuração do Experimento

Seguindo o mesmo procedimento dos experimentos de classificação de notícias falsas, foi necessário avaliar a média do número de sentenças por documento para a definição do número de *frames* e parâmetro K. As notícias apresentam uma média de 21 sentenças por documento. Os artigos, por sua vez, apresentam uma média de 29 sentenças. Apesar dessa maior média, os artigos apresentam muitos textos pequenos, logo foi decidido manter o número de **3 *frames***. Portanto, obtém-se 7 e 9,67 sentenças por *frame*, na média, para notícias e artigos, respectivamente. Além disso, foi mantido o valor de $K = 2$, a fim de evitar que um número significante de artigos não atendesse ao requisito mínimo, necessitando de intenso preenchimento.

Juntamente às decisões discutidas, as demais configurações dos experimentos de classificação de notícias falsas foram mantidas: os modelos de classificação, *Random Forest* e *XGBoost*; os grupos de *features* para avaliação e comparação, AVGM, WMDM, ALFM, SFM0, SFM1 e SFM2; as métricas de avaliação, *Precison, Recall, F1-Score* e *PR-AUC*. Os artigos, por serem a classe minoritária, são considerados a classe positiva. Apesar do número de artigos ser significantemente maior que o de notícias falsas, ainda assim, ele é significantemente menor que o de notícias legítimas, o que levou à manutenção da proporção de 4 notícias para 1 artigo e, também, 500 rodadas de classificação são executadas, havendo randomização dos textos que compõem o conjunto de treino e teste a cada rodada. O mesmo preenchimento do valor da última frase até o tamanho do maior texto presente no 0,9-percentil para o modelo WMDM foi realizado, sendo que, neste caso, o tamanho máximo se manteve em 32 sentenças.

5.3.2 Resultados e Discussão

A Tabela 5.3 apresenta as médias e desvios-padrão dos resultados das métricas *Precision, Recall, F1-Score* e *PR-AUC* das 500 rodadas de cada um dos modelos treinados.

Os resultados apresentados mostram que, usando o mesmo grupo de notícias (legítimas)

Tabela 5.3: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação de artigos de colunas de jornais.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,44 \pm 0,02$	$0,64 \pm 0,02$	$0,75 \pm 0,02$	$0,59 \pm 0,02$	$0,65 \pm 0,02$	$0,73 \pm 0,02$
	XGB	$0,40 \pm 0,33$	$\mathbf{0,93 \pm 0,02}$	$\mathbf{0,89 \pm 0,01}$	$0,78 \pm 0,02$	$0,85 \pm 0,02$	$0,85 \pm 0,02$
Recall	RF	$0,10 \pm 0,01$	$0,15 \pm 0,01$	$\mathbf{0,18 \pm 0,01}$	$0,12 \pm 0,00$	$0,13 \pm 0,01$	$0,18 \pm 0,00$
	XGB	$0,00 \pm 0,00$	$0,10 \pm 0,00$	$\mathbf{0,19 \pm 0,00}$	$0,10 \pm 0,00$	$0,13 \pm 0,00$	$\mathbf{0,20 \pm 0,00}$
F1-Score	RF	$0,17 \pm 0,01$	$0,25 \pm 0,01$	$0,29 \pm 0,01$	$0,20 \pm 0,01$	$0,22 \pm 0,01$	$0,29 \pm 0,01$
	XGB	$0,00 \pm 0,00$	$0,18 \pm 0,01$	$\mathbf{0,32 \pm 0,00}$	$0,17 \pm 0,01$	$0,22 \pm 0,01$	$0,29 \pm 0,00$
PR-AUC	RF	$0,28 \pm 0,01$	$0,37 \pm 0,01$	$0,40 \pm 0,01$	$0,32 \pm 0,01$	$0,34 \pm 0,00$	$0,40 \pm 0,01$
	XGB	$0,32 \pm 0,01$	$0,45 \pm 0,01$	$\mathbf{0,51 \pm 0,00}$	$0,40 \pm 0,00$	$0,43 \pm 0,00$	$0,49 \pm 0,00$

e os mesmos léxicos do experimento de classificação de notícias falsas em português, neste experimento, a eficácia de todos os modelos que se baseiam em fluxo é bem mais satisfatória. Essa observação vem a corroborar com a hipótese de que o desempenho inferior obtido no experimento de notícias falsas em português pode advir da natureza do *dataset* de notícias falsas disponível.

Ao analisar os resultados mostrados pela Tabela 5.3, é possível visualizar que todos os modelos que envolvem as *Audio-Like Features* apresentam melhores resultados que o AVGM em todas as métricas, tanto no *Random Forest* quanto no *XGBoost*. Convém frisar que a eficácia do AVGM no *Recall* e *F1-Score* do *XGBoost* é nula, refletindo no valor da *PR-AUC*. Por sua vez, ao comparar os modelos que envolvem as *Audio-Like Features* e o WMDM, os modelos ALFM e SFM2 obtém melhor eficácia em todas as métricas, exceto na *Precision* do *XGBoost*. Em destaque na tabela, os valores de *Precision* do *XGBoost* apresentados pelo WMDM e pelo ALFM, $0,93$ e $0,89$, respectivamente, sendo ambos mais que o dobro do valor apresentado pelo AVGM. Um outro resultado em destaque é o *Recall* do *XGBoost* apresentado pelo SFM2. Seu valor é $0,20$, levemente superior ao apresentado pelo ALFM ($0,19$), se mostrando 100% maior que o WMDM. Mesmo o WMDM demostrando um ótimo resultado na *Precision* principalmente no *XGBoost*, o baixo valor do *Recall* não permite que a sua eficácia seja maior que os modelos ALFM e SFM2 nas métricas *F1-Score* e *PR-AUC*. Em destaque na Tabela 5.3, o *F1-Score* do *XGBoost* apresentado pelo ALFM, de valor $0,32$, maior $77,78\%$ que o resultado do WMDM. Também em destaque, se apresenta o valor

de *PR-AUC* do *XGBoost* treinado pelo ALFM: $0,51$, sendo $59,37\%$ maior que o AVGM e $13,33\%$ maior que o WMDM. Vale ressaltar novamente que métrica *PR-AUC* é particularmente apropriada para a análise nesses cenários em que há grande desbalanceamento entre as classes, pois, retrata a média dos valores de *Precision* calculados a cada limiar de *Recall*.

Esses resultados mostram que, para esta tarefa, é possível analisar textos usando *features* inspiradas na análise de áudio extraídas de fluxos construídos a partir de distâncias semânticas entre texto e léxico e, ainda, obter bons resultados, respondendo positivamente às questões de pesquisa QP1, QP2, QP3 e QP4. Pela comparação dos resultados dos modelos treinados com *Audio-Like Features* e o AVGM, percebe-se que o uso de *features* extraídas de fluxos apresenta, neste caso, maior eficácia que *features* sintetizadoras com as margens anteriormente citadas, respondendo positivamente às questões QP5 e QP6. O mesmo acontece ao serem comparados os resultados dos modelos quando treinados com as *Audio-Like Features* e o WMDM, respondendo positivamente à QP7, tendo as margens discutidas como resposta à QP8.

Ao analisar apenas os modelos envolvendo *Audio-Like Features* de *frames* individuais, para todas as métricas de ambos os classificadores, a eficácia vai crescendo, em maior ou menor escala, no decorrer dos *frames*, atingindo seu ápice no SFM2, consequentemente. O SFM2 ainda apresenta alguns resultados em igual ou maior patamar que o ALMF. Todos os *Single Frame Models* obtém melhor eficácia que o AVGM. O modelo SFM2 obtém melhor eficácia comparado ao WMDM em todas as métricas, exceto a *Precision* do *XGBoost*. Por sua vez, o SMF1 ainda apresenta maior eficácia que o WMDM na métricas *Precision* do *Random Forest* e *Recall* e *F1-Score* do*XGBoost*. Essas observações sugerem que o aspecto subjetividade é mais decisivo ao diferenciar notícias de artigos de colunas nas porções medial e, principalmente, final do texto. Assim sendo, é possível perceber que o método proposto pode ser capaz de apontar trechos dos textos que apresentam informações mais significativas referentes à tarefa em questão, ao ponto de a análise de apenas um desses trechos superar a eficácia dos modelos de comparação.

Análise de *Audio-Like Features*

A Figura 5.6 apresenta os boxplots da MCR do fluxo de Modalização dos *frames* 0, 1 e 2. Os artigos mostram menores valores de MCR que as notícias em todos os *frames*, porém

maiores medianas. As notícias mantém praticamente os mesmos valores nos dois primeiros *frames*, apresentando uma diminuição no último. Enquanto os artigos apresentam um pico no *frame* 1. Em outras palavras, em relação à dimensão Modalização, as notícias apresentam um comportamento mais estável no início até a porção medial do texto (entre os *frames* 0 e 1); porém apresentam maior variação de subjetividade, nessa dimensão, que os artigos (maior valor de MCR). Ou seja, os autores de artigos tendem mais a manter o mesmo nível de subjetividade em suas frases que os autores de notícias, com um leve aumento dessa variação no *frame* 1.

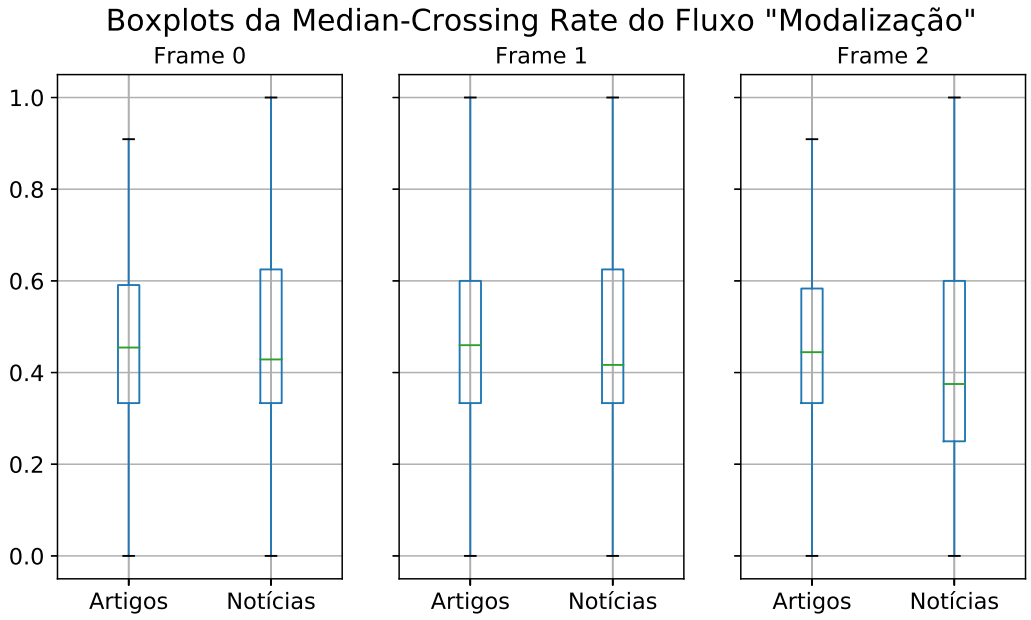

Figura 5.6: Boxplots da *feature Median Crossing Rate (MCR)* do Fluxo "Modalização".

Considerando a dimensão Sentimento, os artigos apresentam menores valores de *Energy Entropy* que as notícias em todos os *frames*, como confirmam os boxplots apresentados na Figura 5.7. Isso leva à percepção de que os artigos sofrem mais mudanças abruptas nas distâncias WMD referentes a esta dimensão. Além disso, os artigos apresentam uma diminuição dos valores no *frame* 1, e um leve crescimento no *frame* 2. As notícias, por sua vez, maném praticamente o mesmo patamar nos dois primeiros *frames*, apresentando diminuição apenas no último. Logo, os artigos apresentam uma maior taxa de mudanças abruptas na porção medial do texto, enquanto as notícias, na porção final.

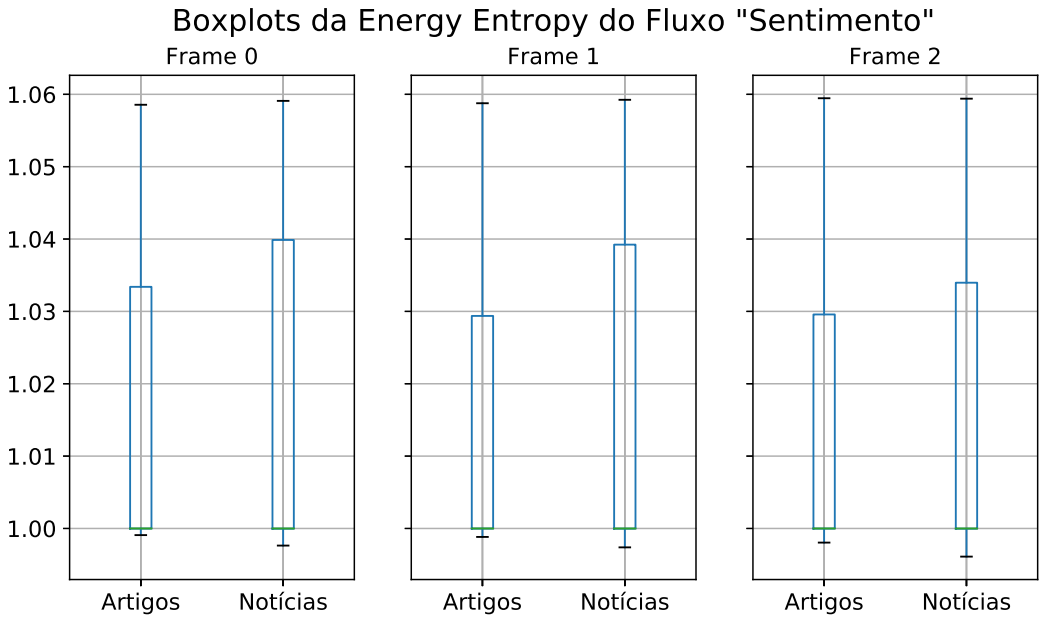

Figura 5.7: Boxplots da *feature Energy Entropy* do Fluxo "Sentimento".

5.4 Classificação de Sentimentos em Avaliações de Filmes

Informações contendo opiniões são amplamente disponíveis *online* e desempenham um papel vital nas avaliações de produtos os serviços, indicando se os consumidores estão satisfeitos ou não. Neste contexto, a análise de sentimentos de avaliações de produtos ou serviços é uma tarefa comumente estudada, dado que foca na classificação de sentimentos (ou opiniões) expressadas em textos gerados por humanos [6].

Com intuito de avaliar o desempenho do método proposto neste âmbito de estudo, foram conduzidos alguns experimentos de classificação de sentimentos em avaliações de filmes escritas em inglês, envolvendo três léxicos muito utilizados para este fim na literatura.

5.4.1 Descrição do Experimento

Datasets

Para os experimentos que serão descritos nesta seção, foi utilizado um *dataset* amplamente difundido em se tratando de avaliações de filmes: o PangLee-04 (PL04) [42].

O PL04 contém 2.000 avaliações de filmes escritas em inglês, sendo 1.000 anotadas como positivas e 1.000 como negativas, todas extraídas do site de avaliações de filmes e programas

de TV IMDb[15].

Léxicos de Sentimento

Com o propósito de construir os fluxos de sentimento (positivo e negativo), foram utilizados, nos experimentos desta seção, três diferentes léxicos de sentimentos utilizados separadamente. Abaixo estão descritos esses léxicos:

- AFFIN [41]: este léxico, em sua mais nova versão até o presente momento, contém um total de 2.477 palavras em inglês, sendo 878 positivas e 1.578 negativas.

- *Bing Liu's lexicon* [26]: este léxico é formado por 2.006 termos positivos e 4.783 negativos, totalizando 6.789 termos. Estes termos incluem, além de palavras relacionadas a sentimentos que são frequentemente utilizadas, palavras escritas incorretamente, gírias e variações comuns. No decorrer do texto, este léxico será tratado por BingLiu.

- SentiWordNet 3.0 [8]: é uma extensão do WordNet[39], um conhecido banco de dados léxico em que palavras são organizadas em uma estrutura de árvore. Consequentemente, cada palavra no SentiWordNet é automaticamente anotada com um valor entre 0 e 1, de acordo com sua positividade ou negatividade. Cada palavra tem um valor associado de ambas as polaridades. Assim como nos experimentos do trabalho de Araque et al. [6], neste trabalho foi calculado um valor agregado de polaridade para representar cada palavra e, assim, rotulá-la como positiva ou negativa. Por exemplo, a palavra "*easy*" possui um valor 0,625 para a polaridade positiva (+0,625) e 0,25 para a negativa (-0,25), resultando numa polaridade de 0,375 ao somar as duas, rotulando a palavra como positiva, portanto. No léxico, as palavras são agrupadas por *synsets*. Como o objetivo é calcular distâncias no espaço de *embedding*, e as palavras ligadas a um mesmo *synset* seriam muito próximas neste espaço, escolheu-se uma palavra representante de cada *synset* para compor a lista de termos ao final. No total, foram obtidos 2.552 termos negativos e 1.173 positivos.

[15]https://www.imdb.com/

Experimento

Neste caso, o objetivo é avaliar o desempenho do método proposto na classificação de sentimentos em avaliações de filmes escritas em inglês.

Inicialmente, cada léxico foi separado em palavras positivas e negativas para serem construídos dois fluxos por texto, o fluxo positivo e o fluxo negativo. A criação dos fluxos segue o padrão dos experimentos anteriormente apresentados: primeiramente, as *stop words* e sentenças menores que 3 palavras são removidas e, então, são calculadas as distâncias WMD de cada sentença das avaliações para as dimensões positiva ou negativa dos léxicos no espaço de *embedding*. A seguir, é procedida a fragmentação dos fluxos em *frames* e são extraídas as *Audio-Like Features*. Como os textos são em inglês, manteve-se o uso do modelo de *word embedding* do Word2Vec, utilizado no experimento de classificação de notícias falsas em inglês.

Configuração do Experimento

Inicialmente, foi realizada a avaliação do número médio de sentenças do *dataset*. O PL04 obteve uma média de 30,99 sentenças por texto. Entretanto, diferentemente dos *datasets* de notícias apresentados neste trabalho, o tamanho dos textos é bem constante, o que permitiu realizar experimentos com **3, 4 e 5 *frames*** usando $K = 2$, sem a necessidade de realização expressiva de preenchimento.

Como modelo de classificação foi escolhido o *Logistic Regression*, posto que este modelo demonstra resultados bem efetivos ao classificar essa tarefa, como afirma o trabalho de Araque et al. [5]. Os grupos de *features* para avaliação e comparação são os mesmos AVGM, WMDM, ALFM, SFM0, SFM1 e SFM2. As métricas de avaliação utilizadas são *Accuracy*, dado o balanceamento entre classes, *Precision, Recall, F1-Score* e *Area Under Receiver Operating Characteristic Curve* (*ROC-AUC*). A *F1-Score* é uma métrica bastante utilizada para avaliar tarefas de classificação de sentimentos em avaliações de filmes [5]. Por sua vez, a *ROC-AUC* é uma métrica apropriada para cenários balanceados como este em questão, que reflete o *trade-off* entre a *True Positive Rate* e a *False Positive Rate* em diferentes limiares de classificação [50; 14]. Logo, é uma melhor métrica de avaliação que a *PR-AUC* para este cenário.

No que diz respeito aos conjuntos de treino e teste, foram seguidas as divisões de *cross-validation* nativamente associadas ao *dataset*. Há 10 divisões, contendo 100 avaliações positivas e 100 negativas, cada. Cada experimento foi executado 10 vezes no total, usando 9 diferentes divisões para treino e a divisão restante para teste a cada execução. Para que esta configuração de cross-validação fosse respeitada ao executar a classificação com o WMDM, o preenchimento foi realizado até o número de sentenças do maior texto do *dataset*.

5.4.2 Resultados e Discussão

Como relatado na Seção 5.4.1, foram executados experimentos usando a divisão em 3, 4 e 5 *frames*. Entretanto, percebeu-se que as diferenças nos resultados eram mínimas, logo, resolveu-se apresentar apenas os resultados com a divisão em 3 *frames*. Nesta seção, apenas serão apresentados os resultados das métricas *F1-Score* e *ROC-AUC*, pois os resultados apresentados por *Precision* e *Recall* são bem equilibrados, neste caso, podendo ser bem representados apenas pela *F1-Score*. Assim, a apresentação desses valores ficará mais clara. Os resultados em sua completude poderão ser conferidos no Apêndice C.

A Tabela 5.4 apresenta os resultados das médias e desvios-padrão dos resultados das métricas *F1-Score* e *ROC-AUC* das 10 rodadas de cada um dos modelos treinados com distâncias calculadas para todos os três léxicos.

Tabela 5.4: Média e desvio-padrão dos resultados das métricas **F1-Score** e **ROC-AUC** da classificação de sentimentos em avaliações de filmes.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Afinn	F1-Score	$0,70 \pm 0,03$	$0,66 \pm 0,02$	$\mathbf{0,72 \pm 0,02}$	$0,62 \pm 0,04$	$0,63 \pm 0,04$	$0,69 \pm 0,03$
	ROC-AUC	$0,70 \pm 0,03$	$0,66 \pm 0,02$	$\mathbf{0,72 \pm 0,01}$	$0,62 \pm 0,03$	$0,63 \pm 0,04$	$\mathbf{0,70 \pm 0,03}$
BingLiu	F1-Score	$0,72 \pm 0,03$	$0,67 \pm 0,03$	$\mathbf{0,73 \pm 0,02}$	$0,63 \pm 0,03$	$0,64 \pm 0,05$	$0,70 \pm 0,03$
	ROC-AUC	$0,73 \pm 0,03$	$0,67 \pm 0,03$	$0,73 \pm 0,02$	$0,63 \pm 0,02$	$0,65 \pm 0,05$	$0,70 \pm 0,03$
SentiWordNet	F1-Score	$0,63 \pm 0,04$	$0,58 \pm 0,05$	$\mathbf{0,66 \pm 0,04}$	$0,56 \pm 0,04$	$0,58 \pm 0,03$	$0,66 \pm 0,04$
	ROC-AUC	$0,66 \pm 0,03$	$0,58 \pm 0,05$	$\mathbf{0,67 \pm 0,04}$	$0,57 \pm 0,03$	$0,58 \pm 0,03$	$0,66 \pm 0,03$

Através desses resultados, é possível perceber que os modelos que envolvem as *Audio-Like Features* apresentam resultados satisfatórios, tendo o ALFM atingido *F1-Score* e *ROC-AUC* de $0,73$ utilizando o léxico BingLiu. Sendo assim, é possível responder afirmativamente às questões de pesquisa QP1, QP2, QP3 e QP4, para este *dataset*, principalmente

quando o léxico BingLiu é usado para criação dos fluxos.

Ainda visualizando a Tabela 5.4, é perceptível que o ALFM obtém resultados iguais ou melhores que o AVGM. Contudo, os resultados de eficácia superior se dão por margens bem menos expressivas que os experimentos anteriores, sendo a maior delas de $4,76\%$, obtida com o *F1-Score* relativo léxico SentiWordNet. Alguns resultados apresentados pelo SFM2 atingem o mesmo valor que o AVGM, como, por exemplo, o *ROC-AUC* relativo ao léxico Afinn. Com isso, ainda assim, a resposta à questão de pesquisa QP5 é afirmativa. Entretanto, as margens, que respondem à QP6, são mais sutis.

Quanto a resposta à questão de pesquisa QP7, os resultados da Tabela 5.4 corroboram para uma resposta afirmativa, uma vez que os modelos ALFM e SFM2 apresentam melhores resultados que o WMDM. Respondendo à QP8, a *ROC-AUC* do ALFM referente ao SentiWordNet, de valor $0,67$, é $13,43\%$ maior que a do WMDM.

Ao analisar apenas os modelos envolvendo *Audio-Like Features*, nenhum dos *Single Frame Models* apresenta melhores resultados que o ALFM. Nos casos referentes a todos os três léxicos, o SFM2 obtém o melhor resultado entre os *Single Frame Models*, mostrando que o modelo é mais eficaz em diferenciar avaliações positivas de negativas na última porção do texto para este *dataset*.

Análise de *Audio-Like Features*

As Figuras 5.8 e 5.9 mostram, respectivamente, os boxplots dos fluxos positivo e negativo gerados a partir do léxico AFINN.

Ao avaliar a Figura5.8, é possível perceber que as avaliações negativas apresentam valores maiores de MCR que as avaliações positivas no primeiro e segundo *frame*, sendo nos terceiro os valores quase iguais. A instabilidade das avaliações negativas quanto à dimensão positiva do léxico decresce ao longo dos *frames*, enquanto as avaliações positivas demonstram um pico no segundo *frame*. Se for observada a Figura 5.9, percebe-se que a situação é exatamente oposta: a instabilidade das avaliações positivas é maior que as negativas e decrescem ao longo do *frame*.

Esse comportamento análogo entre os fluxos das dimensões do léxico e a similaridade que as avaliações positivas e negativas apresentam (representadas por esses boxplots, mas também observadas em outras *features*), sugerem que o comportamento do aspecto senti-

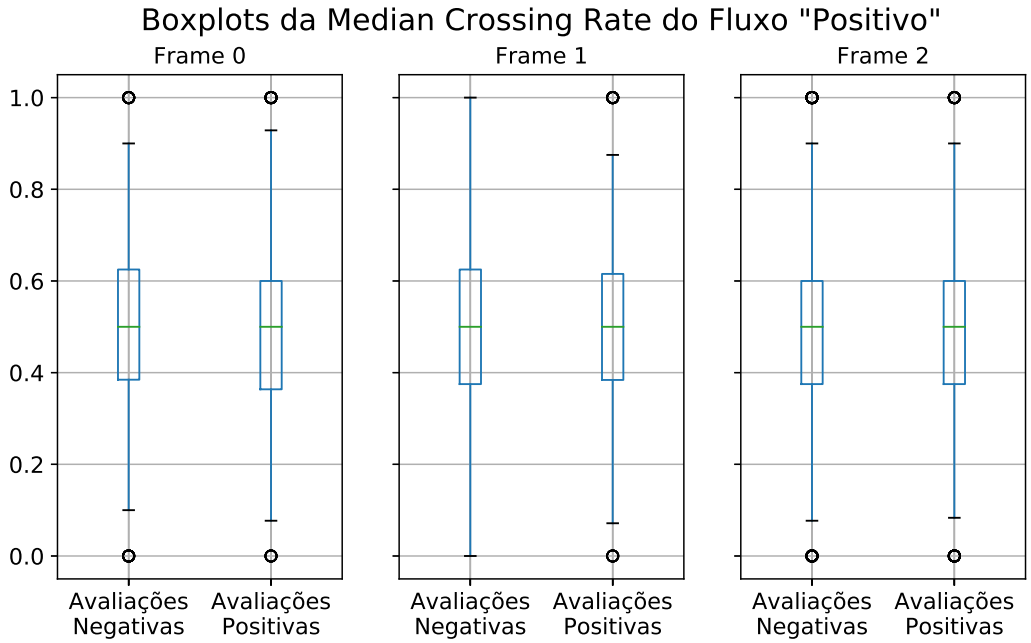

Figura 5.8: Boxplots da *feature Median Crossing Rate (MCR)* do Fluxo "Positivo".

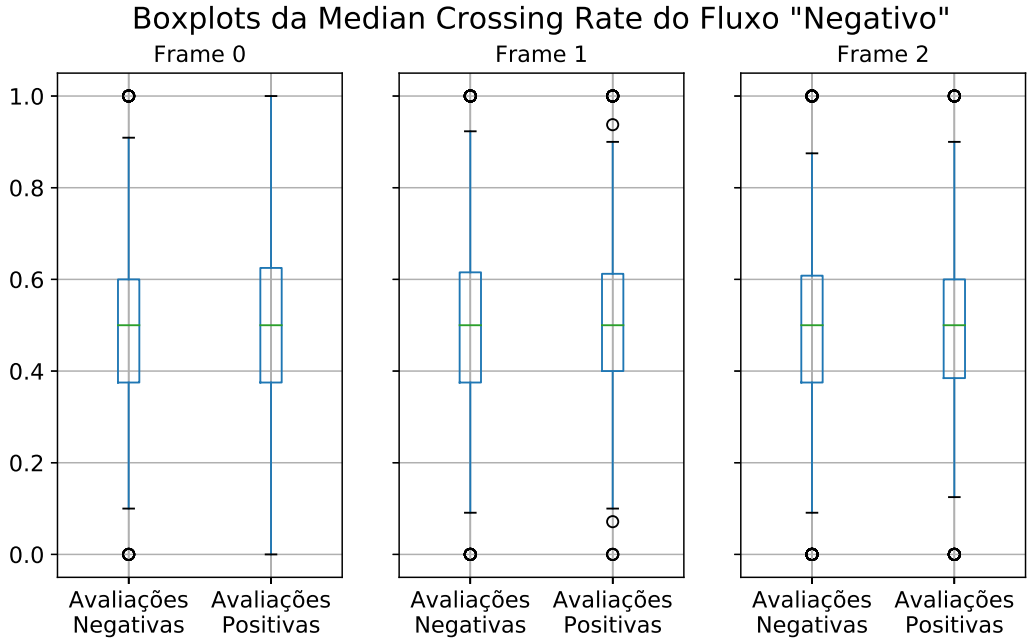

Figura 5.9: Boxplots da *feature Median Crossing Rate (MCR)* do Fluxo "Negativo".

mento é bem neutro nas avaliações de filme, podendo ser esse o motivo pelo qual os resultados obtidos pelo ALFM sejam próximos aos obtidos pelo AVGM.

5.5 Considerações Finais

Neste capítulo, foi apresentada a avaliação experimental do método proposto, abordando três diferentes tarefas de classificação que envolvem aspectos e línguas diferentes.

Através das avaliações dos resultados, é possível perceber que o método demonstra boa eficácia nas tarefas de classificação realizadas. Apenas no experimento de classificação de notícias falsas em português, o método apresenta um menor desempenho, resultado provavelmente devido à natureza do *dataset*, desempenho, contudo, acompanhado pelos métodos usados como comparação, AVGM e WMDM. Mesmo em tarefas com grande desbalanceamento entre classes, o método proposto demonstra bons resultados, superando o AVGM e o WMDM. Na classificação de sentimentos em avaliações de filmes, o método mostra boa eficácia em todas as métricas e supera os modelos de comparação, porém a margem em relação ao AVGM é pequena, o que pode ser explicado pelo comportamento mais neutro do aspecto sentimento nos textos verificado na análise das *features*.

O próximo capítulo apresenta a conclusão do desenvolvimento desta pesquisa de doutorado até o momento e o planejamento das atividades propostas para a conclusão do referido trabalho.

Capítulo 6

Conclusão

A forma como um aspecto linguístico é explorado em um texto pode trazer informações relevantes sobre o mesmo. Para melhor entender como se dá a presença de um certo aspecto, é preciso que a análise seja realizada de forma a capturar como o aspecto se apresenta ao longo de todo o texto. Diferentes tipos de texto podem explorar aspectos de forma distinta no decorrer do texto. Logo, através de uma análise que consiga refletir bem como o aspecto foi explorado desde o início até o fim do texto, é possível capturar as possíveis diferenças entre textos, o que permite dicerní-los.

Portanto, o objetivo dessa pesquisa de doutorado é analisar textos sob a ótica de um determinado aspecto linguístico, preservando a forma como o aspecto é explorado por toda a extensão do texto. Inspirado na semelhança entre o formato dos *Aspect Flows* e de um gráfico de sinal de áudio, o método desenvolvido para analisar os fluxos se baseia na forma como a análise de áudio é realizada, contemplando a divisão dos fluxos em *frames* e a adaptação de *features* que são extraídas a partir deles - as *Audio-Like Features*.

A eficácia do método proposto foi testada em várias tarefas de classificação envolvendo diferentes línguas e aspectos. O artigo *Aspect Flow Representation and Audio Inspired Analysis for Texts* publicada na LREC 2020 [56] (Apêndice A) apresenta alguns dos resultados obtidos quando apenas as *features* de domínio do tempo haviam sido implementadas. Os resultados apresentados nesta proposta incluem as *features* inspiradas no domínio da frequência.

O método se mostra eficaz nas tarefas de classificação propostas, tendo uma eficácia menor no experimento de classificação de notícias falsas em português, resultado provavel-

mente devido à natureza do *dataset*. Mesmo em tarefas com grande desbalanceamento entre classes, o método demonstra bons resultados, superando os modelos usados como comparação, AVGM e WMDM. Na classificação de sentimentos em avaliações de filmes, o método mostra eficácia em todas as métricas e supera os modelos de comparação, porém a margem em relação ao AVGM é pequena, o que pode ser explicado pelo comportamento mais neutro do aspecto sentimento nos textos verificado na análise das *features*.

Diante disso, percebe-se que o método proposto traz resultados promissores, revelando-se como uma forma viável de análise de textos, provendo *features* que podem denotar como um dado aspecto se comporta ao longo dos textos, permitindo a aquisição de conhecimento valioso sobre os mesmos.

6.1 Planejamento

Até a conclusão do desenvolvimento desta pesquisa de Doutorado, ainda se buscam as seguintes contribuições:

- **Contribuição A**: Treinar Redes Neurais Recorrentes, possivelmente redes *Long Short-Term Memory* (LSTM) bi-direcionais, com as *Audio-Like Features* e avaliar o desempenho ao realizar as tarefas de classificação, comparando com os resultados já obtidos e, também, com o desempenho de redes treinadas com o modelo WMDM.

- **Contribuição B**: Avaliar a eficácia do método ao utilizar modelos de *embeddings* que consideram o contexto das palavras na representação, como ELMo [46], BERT [16] ou GPT-2 [47] e métricas compatíveis com esses *word embeddings* [48] para a criação dos *Aspect Flows*.

- **Contribuição C**: Aprimorar o método já desenvolvido através da implementação de novas *features*. Essas novas *features* podem envolver outras *features* inspiradas no domínio de tempo e frequência; *features* que capturam características do domínio do tempo e da frequência ao mesmo tempo, como as propostas por Bhattacharjee et al. [10]; e/ou *features* estatísticas calculadas a partir das *features* de *frames* em sequência (*mid-term features*) [22].

- **Contribuição D**: Implementar uma solução de tratamento de tamanho de fluxos, como, por exemplo, as normalizações sugeridas por Mao e Lebanon [36] ou por Wachsmuth e Stein [58], em busca de uma forma mais suave de obter o número mínimo de sentenças por *frame* para o apropriado cálculo de todas as *features*. Avaliação do impacto que essa solução acarretará nos resultados obtidos nas tarefas de classificação já executadas.

- **Contribuição E**: Avaliar a eficácia do método em tarefas de classificação que envolvam *datasets* contendo textos maiores, permitindo o uso de maior número de *frames* e/ou *frames* mais densos, com mais sentenças.

O planejamento da execução da atividades para a conclusão deste trabalho de doutorado, incluido essas contribuições, está apresentado no cronograma da Tabela 6.1.

Tabela 6.1: Cronograma de atividades.

Atividade	2020			2021							
	10	11	12	01	02	03	04	05	06	07	08
Contribuição A	X										
Contribuição B		X	X								
Contribuição C				X	X						
Contribuição D						X					
Contribuição E							X	X			
Escrita da Tese								X	X	X	
Defesa da Tese											X

Bibliografia

[1] Charu C. Aggarwal and Cheng Xiang Zhai. *Mining Text Data*. Springer Publishing Company, Incorporated, 2012.

[2] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Ahmet Aker, Hauke Gravenkamp, Sabrina Mayer, Marius Hamacher, Anne Smets, Alicia Nti, Johannes Erdmann, Julia Serong, Anna Welpinghus, and Francesco Marchi. Corpus of news articles annotated with article level subjectivity. 06 2019.

[4] Evelin Amorim, Marcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sanchez-Rada, and Carlos A. Iglesias. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236 – 246, 2017.

[6] Oscar Araque, Ganggao Zhu, and Carlos A. Iglesias. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346 – 359, 2019.

[7] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310, 2019.

[8] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[9] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[10] Mrinmoy Bhattacharjee, S. R. Mahadeva Prasanna, and Prithwijit Guha. Time-frequency audio features for speech-music classification. *ArXiv*, abs/1811.01222, 2018.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[12] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36 – 64, 2016.

[13] Yoonjung Choi and Janyce Wiebe. +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar, October 2014. Association for Computational Linguistics.

[14] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[15] Li Deng and Yang Liu. *Deep Learning in Natural Language Processing*. Springer Publishing Company, Incorporated, 1st edition, 2018.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[17] Elena Filatova. Sarcasm detection using sentiment flow shifts. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, pages 264–269, 2017.

[18] Xianghua Fu, Jingying Yang, Jianqiang Li, Min Fang, and Huihui Wang. Lexicon-enhanced lstm with attention for general sentiment analysis. *IEEE Access*, PP:1–1, 10 2018.

[19] Theodoros Giannakopoulos and Aggelos Pikrakis. Chapter 1 - introduction. In Theodoros Giannakopoulos and Aggelos Pikrakis, editors, *Introduction to Audio Analysis*, pages 3 – 8. Academic Press, Oxford, 2014.

[20] Theodoros Giannakopoulos and Aggelos Pikrakis. Chapter 2 - getting familiar with audio signals. In Theodoros Giannakopoulos and Aggelos Pikrakis, editors, *Introduction to Audio Analysis*, pages 9 – 31. Academic Press, Oxford, 2014.

[21] Theodoros Giannakopoulos and Aggelos Pikrakis. Chapter 3 - signal transforms and filtering essentials. In Theodoros Giannakopoulos and Aggelos Pikrakis, editors, *Introduction to Audio Analysis*, pages 33 – 57. Academic Press, Oxford, 2014.

[22] Theodoros Giannakopoulos and Aggelos Pikrakis. Chapter 4 - audio features. In Theodoros Giannakopoulos and Aggelos Pikrakis, editors, *Introduction to Audio Analysis*, pages 59 – 103. Academic Press, Oxford, 2014.

[23] Yoav Goldberg and Graeme Hirst. *Neural Network Methods in Natural Language Processing*. Morgan Claypool Publishers, 2017.

[24] Patrick Helmholz, Michael Meyer, and Susanne Robra-Bissantz. Feel the moosic: Emotion-based music selection and recommendation. 06 2019.

[25] Benjamin D. Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, 2017.

[26] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA, 2004. Association for Computing Machinery.

[27] Madiha Jalil, Faran Butt, and Ahmed Malik. Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. pages 208–212, 05 2013.

[28] Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence & Applications*, 6, 06 2015.

[29] Caio Jeronimo, Leandro Marinho, Claudio Campelo, Adriano Veloso, and Allan Melo. Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, 2019.

[30] Lakshmish Kaushik, Abhijeet Sangwan, and John Hansen. Sentiment extraction from natural audio streams. 05 2013.

[31] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org, 2015.

[32] Caio L. M. Jeronimo, Claudio E. C. Campelo, Leandro Balby Marinho, Allan Sales, Adriano Veloso, and Roberta Viola. Computing with subjectivity lexicons. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3272–3280, Marseille, France, May 2020. European Language Resources Association.

[33] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[34] Seung-Wook Lee, Jung-Tae Lee, Young-In Song, and Hae-Chang Rim. High precision opinion retrieval using sentiment-relevance flows. pages 817–818, 01 2010.

[35] Lie Lu and Hao Jiang. Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10:504 – 516, 11 2002.

[36] Yi Mao and Guy Lebanon. Isotonic conditional random fields and local sentiment flow. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 961–968. MIT Press, 2007.

[37] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[39] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[40] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250, 2012.

[41] Finn Aruprup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[42] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.

[43] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135, January 2008.

[44] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[45] Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA, October 2010. Association for Computational Linguistics.

[46] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[47] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[48] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. Enhancing unsupervised sentence similarity methods with deep contextualised word representations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 994–1003, Varna, Bulgaria, September 2019. INCOMA Ltd.

[49] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[50] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. In *PloS one*, 2015.

[51] Jangwon Seo and Jiwoon Jeon. High precision retrieval using relevance-flow graph. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 694–695, New York, NY, USA, 2009. ACM.

[52] C Silverman. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. buzzfeed, nov. 16, 2016.

[53] Roberta Sinoara, José Camacho-Collados, Rafael Rossi, Roberto Navigli, and Solange Rezende. Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 10 2018.

[54] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[55] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[56] Larissa Vasconcelos, Claudio Campelo, and Caio Jeronimo. Aspect flow representation and audio inspired analysis for texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1469–1477, Marseille, France, May 2020. European Language Resources Association.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[58] Henning Wachsmuth and Benno Stein. A universal model for discourse-level argumentation analysis. *ACM Trans. Internet Technol.*, 17(3):28:1–28:24, June 2017.

[59] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September 2004.

[60] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 347–354, USA, 2005. Association for Computational Linguistics.

[61] Chuhan Wu, Fangzhao Wu, Junxin Liu, Yongfeng Huang, and Xing Xie. Sentiment lexicon enhanced neural sentiment classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1091–1100, New York, NY, USA, 2019. Association for Computing Machinery.

Apêndice A

Aspect Flow Representation and Audio Inspired Analysis for Texts

Aspect Flow Representation and Audio Inspired Analysis for Texts

Larissa L. Vasconcelos[1,2], Claudio E. C. Campelo[1], Caio L. M. Jeronimo[1]
Federal University of Campina Grande[1], Federal Institute of Paraiba[2]
882, Aprigio Veloso, Campina Grande - PB - Brazil[1]
PB-264, Monteiro - PB - Brazil[2]
larissa.vasconcelos@ifpb.edu.br, campelo@computacao.ufcg.edu.br, caiolibanio@copin.ufcg.edu.br

Abstract

For better understanding how people write texts, it is fundamental to examine how a particular aspect (e.g., subjectivity, sentiment, argumentation) is exploited in a text. Analysing such an aspect of a text as a whole (i.e., through a summarised single feature) can lead to significant information loss. In this paper, we propose a novel method of representing and analysing texts that consider how an aspect behaves throughout the text. We represent the texts by aspect flows for capturing all the aspect behaviour. Then, inspired by the resemblance between these flows format and a sound waveform, we fragment them into frames and calculate an adaptation of audio analysis features, named here Audio-Like Features, as a way of analysing the texts. The results of the conducted classification tasks reveal that our approach can surpass methods based on summarised features. We also show that a detailed examination of the Audio-Like Features can lead to a more profound knowledge about the represented texts.

Keywords: Text Representation, Text Analysis, Aspect Flows

1. Introduction

Natural Language Processing (NLP) is a branch of Artificial Intelligence that tries to understand how humans communicate. An NLP challenge is to represent human language in a manner that permits the extraction of valuable information aiming at exploiting its singularities.

For a better understanding of how different kinds of texts are written, it is important to evaluate how they exploit a particular aspect. For instance, analysing how fake news make use of subjectivity can lead to meaningful knowledge about them (Jeronimo et al., 2019). As an aspect, we consider distinct types of information that a text may contain, such as sentiment, argumentation, or subjectivity. Observing the particularities of how a text utilises some aspect requires not only a global analysis of the entire text or individual analysis of each word or sentence, but also an analysis of its behaviour throughout the text.

Generally, for executing NLP tasks, the applied techniques model local representations of an aspect to extract summarised features that represent the input text as a whole, frequently by an average or median of that local representations. This sort of representation can lead to relevant information loss, especially for large texts, as they can be ignoring significant aspect singularities present in any part of the text that could be decisive on text type identification and characterisation.

An approach to avoiding representing a text in a globally summarised way is using flows, which can be defined as a sequence of information collected from the words, sentences, or paragraphs of the text. Mao and Lebanon (2007) use sentiment flows to represent texts by assigning for each sentence one of the following values: 2 (highly praised), 1 (something good), 0 (objective description), -1 (something that needs improvement) and -2 (strong aversion). They propose a variant of conditional random fields (Lafferty et al., 2001) to proceed local and global sentiment prediction in reviews. Wachsmuth and Stein (2017) represent the text's discourse-level structure as a flow of rhetorical moves. They model until four kinds of text segment flow: local sentiment, modeling negative, neutral, and positive sentiment; discourse relation between segments, e.g., cause, circumstance, condition; paragraph-level discourse functions (introduction, body, rebuttal, conclusion) (Persing et al., 2010); and argument roles, modeling real arguments, premises or claims. They propose a clusterisation in training flows and compare test flows to the training cluster's centroids in order to perform global reviews sentiment classification and essay scoring. Filatova (2017) models product reviews as sentiment flows and uses sentiment changing for sarcasm detection. She uses the Stanford Sentiment Analysis tool (Socher et al., 2013) with the 5-point sentiment scale (very negative (-2), negative (-1), neutral (0), positive (+1), very positive (+2)) to assign sentiment labels to each sentence in texts. Lee et al. (2010) represent texts as a merge of a sentiment flow and a relevance flow (Seo and Jeon, 2009) to proceed with opinion retrieval. For each sentence of the text, they calculate a score that reflects its relevance (concerning a query) and opinion (the frequency of a lexicon's opinion words). As features, they use the variance of sentence scores, the fraction of peaks, and the first peak position.

In this paper, we propose to represent texts as aspect flows and perform a sophisticated flow analysis based on the concept of frame from audio analysis. The solution is independent of the target aspect being investigated, so its selection depends on what kind of information is meaningful to a given NLP task. In order to obtain an informative manner of analysing the way the aspect behaves throughout the text, our proposed approach divides the aspect flows into frames. Then, it extracts the so-called Audio-Like Features, an adaptation of audio analysis features for the text-domain. We evaluated the model in three NLP classification tasks: Fake News Classification Based on Text Subjectivity, Newspaper Columns Classification Based on Text Subjec-

tivity, and Movie Reviews Sentiment Classification. The first task uses subjectivity flows to explore the differences between legitimate and fake news, since fake news is more subjective than legitimate news (Jeronimo et al., 2019). In the second task, we also generate subjectivity flows, however, the aim is to differentiate objective news from newspaper columns, since newspaper columns are opinionated texts and objective news should not be. The latter task, on the other hand, investigates the overall movie reviews sentiment by analysing flows made of sentiment polarities.

Through the evaluation tasks, the proposed model reveals to be a viable form to represent and analyse texts, providing meaningful features for examining how a given aspect behaves throughout the texts, making it feasible to acquire valuable knowledge about the subject tasks.

The rest of the paper is structured as follows. Section 2. describes the proposed text representing and analysing model. Following, in Section 3., we detailed explore the executed experiments, including used datasets, lexicons, and experimental setup, as well as these experimental results and discussion. Finally, the paper concludes with Section 4., which depicts the conclusions drawn from the evaluation and outlines the possible future lines of work.

2. Proposed Model

This section introduces our proposed model, describing the aspect flows generation, the division of flows into frames, and the subsequent extraction of the features inspired by audio analysis, the audio-like features. Figure 1 shows a diagram of the proposed model.

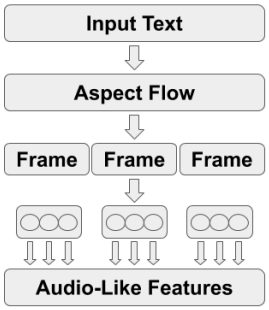

Figure 1: Model Diagram.

2.1. Aspect Flows Generation

Representing texts by a flow of an aspect related to a task is a promising way to better understand how the aspect behaves in the text, as the sentiment flow modeling explored in Mao and Lebanon (2007) article confirms.

For generating the aspect flow representation of the text, first it is necessary to split the text into sentences and then obtain an aspect representation for each one. The flow is the sequence of these sentence aspect representations. The aspect representation of a sentence can be generated, for instance, through a model trained on an annotated dataset, or via a model based on semantic similarity computation between the sentence and an aspect lexicon. All three tasks performed in this paper use the latter method to construct the flows, making annotated bases not necessary.

2.2. Audio-Like Features Extraction

If we plot an aspect flow as a graphic using the x-axis to represent the sentences and the y-axis the aspect values, we can perceive similarities between the form of this plot and a graphic of a sound waveform. In order to illustrate that, Figure 2 shows an example of an argumentation subjectivity flow of a legitimate news in our dataset. Given the

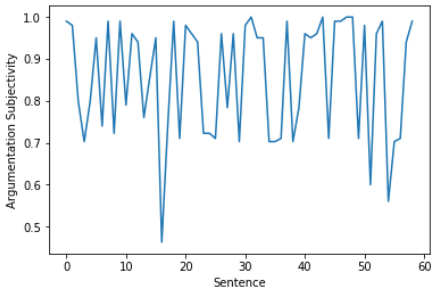

Figure 2: Example of an Argumentation Subjectivity Flow Graphic.

above, we propose to adapt the manner audio analysis is performed to the text's domain. Audio analysis is a solidified research area focused on providing useful knowledge about audio content (Giannakopoulos and Pikrakis, 2014a). This knowledge has proved valuable in many fields, such as segmentation and classification for music recommendation (Lu and Jiang, 2002; Helmholz et al., 2019), speech-music classification (Bhattacharjee et al., 2018), song emotion analysis (Jamdar et al., 2015), and sentiment extraction from speech streams (Kaushik et al., 2013).

2.2.1. Aspect Flow Frames

Frequently, in audio analysis, the audio signal is broken into possibly overlapping short-term windows (frames), and the analysis is carried out on a frame basis. In order to understand the reason for using this windowing technique, consider one audio which presents a conversation and a gunshot in the middle of that. If we compute an average intensity of the samples of the whole recording, the samples presenting the gunshot will dominate the result. If we analyse just this metric, we can obtain disturbed conclusions about the audio. Hence, it seems more feasible to compute features from the audio frames to better represent the information present there (Giannakopoulos and Pikrakis, 2014b).

As we aim to examine how aspects behave throughout texts, we propose to adopt the short-term windowing technique, fragmenting aspect flows into, initially, non-overlapping frames. In order to be able to compare the same parts of different texts (which very often have different sizes), our model breaks the aspect flows in a fixed number of frames. Therefore, regardless of the number of sentences in a text, the first frame will represent the first part of the text, for example, which we can compare to another text's first

part. Concerning defining the number of frames to split the flows, it is a dataset-dependent decision, as, for instance, if we are dealing with books, we can obtain so much more meaningful frames than with movie reviews.

2.2.2. Aspect Flow Audio-Like Features

In audio analysis, there are two categories of frame extracted features: time-domain and frequency-domain. The time-domain features offer a simple way to analyse audio signals and are directly extracted from the samples of the audio signal (waveform). On the other hand, frequency-domain features are extracted from the sound spectrum, a representation of the distribution of the frequency content of sounds (Giannakopoulos and Pikrakis, 2014d). Obtaining this representation requires to compute the Discrete Fourier Transform (DFT) of the audio signal (Giannakopoulos and Pikrakis, 2014c).

Our model presents the Audio-Like Features, an adaptation of audio analysis features extracted from the texts' aspect flows. Initially, it implements only the time-domain feature extraction, since it is possible to perform it directly from the flows. Our three Audio-Like Features are Energy, Median-Crossing Rate, and Energy Entropy, and will be detailed hereafter.

The first Audio-Like Feature is Energy. As the original version, this feature reflects the total magnitude of the aspect in the flow (Jalil et al., 2013). Let $x_i(n), n = 1, ..., F_L$ be the sequence of sentences of the i-th aspect frame, where F_L is the length of the frame. The implementation of Energy is defined as:

$$E(i) = \frac{1}{F_L} \sum_{n=1}^{F_L} |x_i(n)|^2 \tag{1}$$

Here we normalised the Energy by dividing it by F_L to remove the dependency on the frame length. The stronger an aspect appears in the frame, the bigger the frame's Energy.

Median-Crossing Rate (MCR) is the adaption of audio frame feature Zero-Crossing Rate (ZCR), which is the rate of sign-changes of the signal during the frame. As the audio signal waveform amplitude varies from -1 to 1, the ZCR is the number of times the signal changes value, from positive to negative and vice versa, divided by the length of the frame [livro cap4]. As we cannot expect all the sort of aspect inputs to be in the same audio signal amplitude's range $[-1, 1]$, our implementation uses the aspect flow median ($flowmedian$) as the parameter to calculate the crossing rate. The MCR is defined according to the following equation:

$$MCR(i) = \frac{1}{2F_L} \sum_{n=1}^{F_L} |msgn[x_i(n)] - msgn[x_i(n-1)]| \tag{2}$$

where $msgn$ is a modification of sign function, the Median Sign Function, denoted by:

$$msgn[x_i(n)] = \begin{cases} 1, & \text{if } x_i(n) > flowmedian. \\ -1, & \text{if } x_i(n) < flowmedian. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

MCR can be interpreted as a measure of the target aspect noisiness of a flow; in other words, it reflects the level of aspect variation in a flow frame.

The last aspect feature is the Entropy of Energy, which can be interpreted as a measure of abrupt changes in the energy level of an aspect flow (like in audio analysis (Giannakopoulos and Pikrakis, 2014d)). For example, it can detect if a frame presents sentences with deeply different levels of subjectivity. In order to extract it, we first divide each flow frame into K sub-frames. This parameter must not be less than $K = 2$ to ensure there will be at least 2 sub-frames. Hence, it depends on the mean frame size of the aspect flows generated from a dataset, so that the model can avoid the unwanted effect of generating various sub-frames in a short frame. Then, for each sub-frame j, we compute its Energy as in (1) and divide it by the total frame Energy, $Eframe_i$. The division is necessary to treat the resulting sequence of sub-frame energy values, $e_j, j = 1, ..., K$, as a sequence of probabilities, as in (4):

$$e_j = \frac{Esubframe_j}{Eframe_i} \tag{4}$$

where

$$Eframe_i = \sum_{k=1}^{K} Esubframe_k \tag{5}$$

At a final step, the entropy, $Ent(i)$ of the sequence e_j is computed according to the equation:

$$Ent(i) = -\sum_{j=1}^{K} e_j * log_2(e_j) \tag{6}$$

The more significant changes the frame presents, the lower the Entropy Energy resulting value is.

As we can notice in Equations (2) and (6), to calculate MCR and Energy Entropy correctly, the flow must contain, at least, 2 sentences per frame. Thus, this minimum requirement must be considered during the definition of the number of frames and the K parameter (as it requires, at least, one sentence per subframe). Considering this, our model is not appropriate to analyse tiny texts, such as those microblogs.

3. Experimental Evaluation and Discussion

In this section, we describe our experimental evaluation conduction, also presenting the results of the three evaluation tasks and a discussion about these obtained results. For each task, we describe the utilised datasets, the linguistics resources used to generate the aspect flows, the evaluation protocol, the performance measures, and the classification models.

3.1. Fake News Classification Based on Text Subjectivity

The need for fake news detection is clear and present given the ubiquitous reach of social media sites like Facebook and Twitter. Recent events indicate the growing power of the fake news dissemination as 2018 Brazilian presidential elections, in which there is much evidence about massive dissemination of misleading content by the illegal usage of messaging applications[1].

[1] https://www.dw.com/en/brazil-police-to-probe-allegations-of-electiondisinformation-on-whatsapp/a-45965369

Jeronimo et al. (2019) demonstrated good results in performing Fake News Classification Task in a dataset of Brazilian legitimate and fake news, considering that their subjectivity levels are significantly different. For that, the authors rely on a set of subjectivity lexicons built by Brazilian linguists (Amorim et al., 2018) and build subjectivity feature vectors for each news article. For generating these feature vectors, the Word Mover's Distance (WMD) (Huang et al., 2016) between each news sentences and these lexicons considering the embedding the news words lie in is calculated. Then, an average of the distances of each document sentences to each lexicon is computed. As Jeronimo et al. (2019) use a summarised way to represent text aspects (average of sentences' WMD), we decided to evaluate our model by replicating their most challenging experimental scenario, which consists of fake news classification regardless of the domain and sources of legitimate and fake news.

3.1.1. Dataset

In this paper, we use the same dataset as Jeronimo et al. (2019). The dataset of legitimate news was collected from two of the biggest news sites in Brazil, that are Estadao[2] and Folha de Sao Paulo[3]. The dataset has a total of 207,914 legitimate news, from the years 2014 to 2017, divided into different domains: Politics, Sports, Economy, and Culture. The fake news dataset is composed of fact-checked fake news that strongly disseminates in Brazil, from the years of 2010 to 2017. These news were collected from two popular fact-checking services, that are e-Farsas[4] and Boatos[5]. The fake news dataset is formed by a total of 121 fake news from more than 40 news sources.

3.1.2. Subjectivity Lexicons

As the source of aspect information, we employ the same five Brazilian Portuguese subjectivity lexicons (Amorim et al., 2018) Jeronimo et al. (2019) did. These lexicons were built by Brazilian linguists and are described next:

- The argumentation dimension represents words and expressions that are related to a more argumentative discourse. Such discourse is often used when someone is trying to convince another person of a specific point of view.
- The presupposition dimension encompasses terms that are related to a previous assumption of something. This kind of discourse is mainly used in situations where the interlocutor assumes something as true, even when this is not the case.
- The sentiment lexicon contains words and terms related to emotional discourse. Such terms are also used in the context of fake news when the writer of the article tries to emotionally engage the reader.
- The valuation dimension expresses words related to the amount or intensification of something.
- The modalization discourse is used when the interlocutor has an established stance about something or someone.

3.1.3. Experiment

The main objective of the experiment is to evaluate how effective are the subjectivity flow representation and Audio-Like Features analysis for fake news classification. For the purpose of building the subjectivity flows, we use the same method as Jeronimo et al. (2019): for each news, we calculate its sentences' WMD to the five subjectivity lexicons considering the embedding the news words lie in. The main difference is that, instead of using an average of these WMD values, we use the sequence of them as an aspect flow to represent the news. Therefore, for each news in the dataset, we generate five subjectivity flows, one to each lexicon. Then, we fragment each flow in frames and calculate the three Audio-Like Features for every frame of the flow. We use all legitimate news, regardless of the domain and sources, and the fact-checked fake news as model input data. This is a challenging scenario because the legitimate and fake news are a mix of different domains and sources.

3.1.4. Experimental Setup

We evaluate the average number of sentences per document in the dataset to define the number of frames the subjectivity flows should be split into, and the value that should be assigned to the K parameter to calculate the Energy Entropy. Legitimate and fake news contain an average of 21 and 14 sentences per document, respectively. Thus, we decided to split the flows into 3 frames, resulting into 7 and 4.67 sentences per frame, on average, for the legitimate and fake news, respectively. Since we have obtained frames with few sentences on average, we decided to use $K = 2$, to have at least two sentences per sub-frame, on average, the minimal necessary number to calculate the Energy Entropy correctly. Considering this decision, many documents did not meet the minimum requirement, and then we had to discard them. We then performed the classification task with 187,194 legitimate news and 88 fake news, less than Jeronimo et al. (2019) experiment.

To evaluate the applicability of our proposed features for classifying fake news and to compare our results to those of Jeronimo et al. (2019), we used the model that best performed in their experiments: Random Forest. As the dataset of legitimate news is far more significant than the fake one, we randomize the train/test executions by varying the legitimate news documents 500 times. We also follow the proportion of four legitimate news to one fake news (Silverman, 2016). To calculate the semantic distances with WMD, we used the word embedding model from a large Wikipedia dump in Portuguese trained by Jeronimo et al. (2019). We evaluate the models in terms of the Area Under the Precision-Recall curve (PR-AUC), a metric that suits our scenario of class imbalance (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006).

3.1.5. Results and Discussion

We performed the classification task by training Random Forest models with: (1) the Jeronimo et al. (2019) Average Features (Average Model); (2) the Audio-Like Fea-

[2]https://www.estadao.com.br/
[3]https://www.folha.uol.com.br/
[4]http://www.e-farsas.com/
[5]http://www.boatos.org/

tures of all three frames (All Frames Model); and (3) the Audio-Like Features per individual frame (Single Frame Model). The average of the PR-AUC for all 500 runs for each trained model is shown in Table 1. It is possible to visualize that the All Frames Model outperformed the Average Model by 39%. Moreover, it can be noticed that all Single Frame Models obtained better results than the Average Model, especially the Single Frame 1 Model that outperformed it by $\approx 42\%$. This model even performed better than All Frames Model, showing that the subjectivity aspect is more decisive in differentiating legitimate and fake news in the middle of the texts. These results show that representing texts with aspect flows and analysing them using Audio-Like Features can improve the power of classification and possibly point the text's excerpt that presents more meaningful information referring to the task.

3.1.6. Audio-Like Features Analysis

We proceed with some analysis of the Audio-Like Features throughout the three frames, to exemplify what kind of information about the dataset our method can provide. Figure 3 shows the MCR boxplots of the presupposition flow for frames 0, 1 and 2. It can be seen that fake news values are higher than legitimate ones in all three frames. In addition, fake news MCR values increase throughout the frames, while legitimate news MCR values remain stable. These findings show that fake news is more unstable related to the presupposition lexicon distances than legitimate news, mainly in the last portion of the text. The boxplots of the Energy Entropy for the argumentation flow are shown in Figure 4. It can be observed that fake news present higher values than legitimate news in the first frame, and then this situation changes in the other frames. This information means that, in the beginning of the text, legitimate news undergo more abrupt changes in the WMD to the argumentation lexicon, whereas such abrupt changes occur in the middle and in end of the text in the case of fake news.

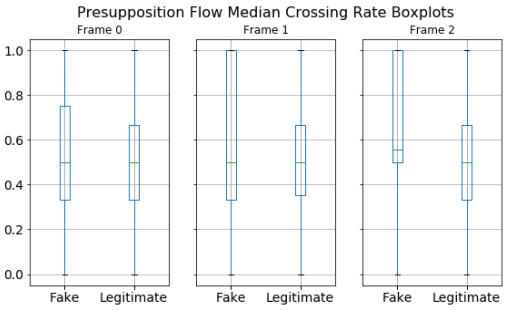

Figure 3: Boxplots of the Median Crossing Rate (MCR) for the Presupposition Flow.

3.2. Newspaper Columns Classification Based on Text Subjectivity

Column is a recurring feature written by the same author in a newspaper. It is often characterised by the voice, personality, and opinions of the writer, in opposition to objective news that reports facts. Considering these characteristics, it seems feasible to discriminate between objective news and newspaper columns based on subjectivity of the language used in texts. Hence, we evaluate the task of classifying Newspaper Columns Classification Based on Text Subjectivity, by adopting a methodology similar to that presented in Section 3.1., based on the WMD to subjectivity lexicons.

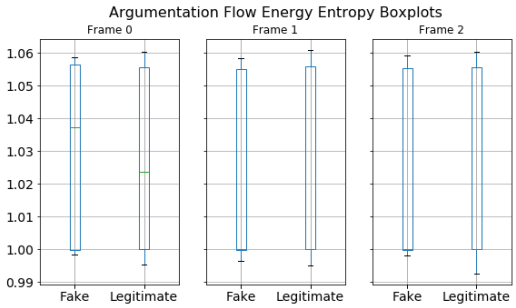

Figure 4: Boxplots of the Energy Entropy for the Argumentation Flow.

3.2.1. Dataset

We use the legitimate news dataset, presented in Section 3.1.1., to represent the objective news, as they are not collected from columns. The newspaper columns dataset was collected from Folha de Sao Paulo[6]. We collected a total of 7,062 columns articles, from a variety of domains (e.g., politics, economy, business, tourism) from 2010 to 2018.

3.2.2. Experiment

The main objective of this experiment is to evaluate how effective are the subjectivity flow representation and the Audio-Like Features analysis for objective news versus newspaper columns classification. The flows generation and Audio-Like Features calculation follow the same steps described in Section 3.1.3.. We use all the objective news and columns articles, regardless of the domain and sources, to keep the challenging character of the task.

3.2.3. Experimental Setup

As above mentioned, the objective news presents an average of 21 sentences per document, but the newspaper columns dataset presents an average of 29 sentences. Despite having a more significant number of sentences average, the newspaper columns dataset present lots of smaller texts, so, to prevent dropping a notable number of texts, we decided to maintain the number of 3 frames; therefore, we have 7 and 9.67 sentences per frame, on average, for the objective news and newspaper columns dataset, respectively. Additionally, we have maintained $K = 2$ to ensure there will be at least three sentences per sub-frame, in order to evaluate the model using bigger sub-frames to calculate the Energy Entropy. In spite of this, a lot of documents do not yet meet the minimal requirements, and therefore we had to discard them, reducing the dataset used in this experiment to 187,194 objective news and 5,429 column articles.

[6]https://www.folha.uol.com.br/

	AVG Features	AL Features All Frames	AL Features Frame 0	AL Features Frame 1	AL Features Frame 2
PR-AUC	0.25 ± 0.03	0.41 ± 0.05	0.29 ± 0.04	0.43 ± 0.06	0.29 ± 0.04

Table 1: Average PR-AUC results for the models trained with Jeronimo et al. Average Features (AVG Features), Audio-Like Features of all frames (AL Features All Frames) and Audio-Like Features per frame (AL Features Frame $\{0,1,2\}$).

We also kept the other setup guidelines,such as the use of the Random Forest model for evaluating the applicability of our model compared to the Average Features. Although the newspaper columns dataset is more significant than the fake news dataset, it is still far less significant than the objective one, therefore we keep the 500 times randomisation and the four to one proportion. The Area Under Precision-Recall curve (PR-AUC) remains the metric used to evaluate the models.

3.2.4. Results and Discussion

The average PR-AUC for all 500 runs of each trained model is shown in Table 2. Once more, the All Frames Model outperformed the Average Model (now the difference was of $\approx 23\%$). All Single Frame Models also performed better than the Average Model. The Single Frame 2 Model has surpassed the Average Model at $\approx 19\%$. Although in this experiment no Single Frame Model has outperformed the All Frames Model, the better performance of Single Frame 2 Model indicates that the subjectivity aspect is more efficient in discerning objective news from newspaper column in the ending excerpt of the texts. These results show that our proposed method can potentially improve the classification achievements, also pointing the most discriminative excerpt of the texts.

3.2.5. Audio-Like Features Analysis

Figure 5 shows the MCR boxplots of the sentiment flow for the 3 frames. Newspaper columns present smaller MCR values than objective news over all frames. Objective news maintains almost the same values throughout the frames, while column news presents a peak in the second frame. In other words, regarding the sentiment lexicon, the objective news is more unstable throughout the text, and newspaper columns present more instability in the middle of the text. Newspaper columns show smaller energy entropy values than objective news regarding the modalization flow in all frames, as the boxplots presented in Figure 6 confirm. From these analysis, we can conclude that newspaper columns undergo more abrupt changes in the WMD to the referred lexicon.

3.3. Movie Reviews Sentiment Classification

Opinionated information is widely available online and plays a vital role in evaluating whether a product or service is pleasing their consumers or not. In this context, sentiment analysis of product or service reviews is a common exploited field since it focuses on the classification of sentiments or opinions expressed in human-generated texts (Araque et al., 2019). In order to evaluate our proposed method performance against a summarised approach in several domains of texts and in different languages, we also conduct an experiment on sentiment classification of movie

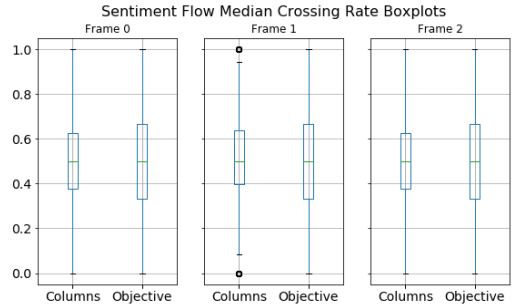

Figure 5: Boxplots of the Median Crossing Rate (MCR) for the Sentiment Flow.

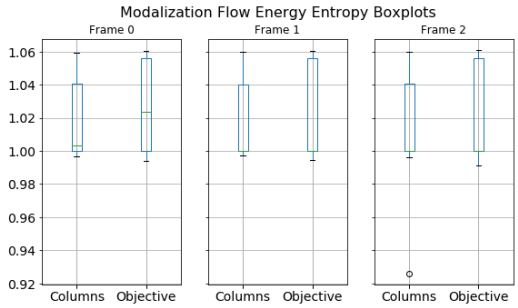

Figure 6: Boxplots of the Energy Entropy for the Modalization Flow.

reviews written in English.

3.3.1. Dataset

The dataset used in this task is the PL04 (Pang and Lee, 2004), containing 2,000 movie reviews written in English. There are 1,000 positive and 1,000 negative labeled movie reviews extracted from the IMDb[7] site of movies and TV shows reviews.

3.3.2. Sentiment Lexicon

With the purpose of constructing the sentiment flows, we calculate the WMD of the sentences of the PL04 texts to the sentiment lexicon AFINN (Nielsen, 2011). We used the newest version of AFINN, that contains a total of 2,477 English words, being 878 positives, and 1,578 negatives. Positive words are scored from 1 to 5, while negative ones have a sentiment score ranging from -5 to -1.

[7] https://www.imdb.com/

	AVG Features	AL Features All Frames	AL Features Frame 0	AL Features Frame 1	AL Features Frame 2
PR-AUC	0.34 ± 0.01	0.44 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	0.42 ± 0.01

Table 2: Average PR-AUC results for the models trained with Jeronimo et al. Average Features (AVG Features), Audio-Like Features of all frames (AL Features All Frames) and Audio-Like Features per frame (AL Features Frame $\{0,1,2\}$).

	AVG Features	AL Features All Frames	AL Features Frame 0	AL Features Frame 1	AL Features Frame 2
F1 score	0.70 ± 0.03	0.72 ± 0.01	0.62 ± 0.04	0.64 ± 0.05	0.69 ± 0.03

Table 3: Average F1 score results for the models trained with Jeronimo et al. Average Features (AVG Features), Audio-Like Features of all frames (AL Features All Frames) and Audio-Like Features per frame (AL Features Frame $\{0,1,2\}$).

3.3.3. Experiment

In this case, we want to evaluate how beneficial are sentiment flow representation and Audio-Like Feature analysis for movie reviews sentiment classification. First, we separate the AFINN's negative from positive words, generating two polarity lexicons. Then we generate two sentiment flows (negative and positive) for each review, calculating the WMD to the lexicons. Afterwards, we proceed with the flow fragmentation into frames and the Audio-Like Features calculation, as performed in the other experiments.

3.3.4. Experimental Setup

We use the Logistic Regression model for evaluating the applicability of our approach, as this model shows quite effective results in executing this task (Araque et al., 2017). Concerning the training and test procedures, we have followed the PL04 associated cross-validation splits, which is composed of 10 splits, with 100 positive and 100 negative reviews each. We performed 10 executions in total, using 9 different splits to train and the remainder split to test the model in each execution. To calculate the semantic distances with WMD, we used the widely widespread pre-trained word vectors of Word2Vec approach[8]. To evaluate the models, we use the F1 score, a metric that seeks a balance between Precision and Recall (Araque et al., 2017). The length of reviews is of 30 sentences on average. However, this dataset contains several smaller texts, that will be dropped if we fragment them into more than 3 frames. For this reason, we kept the division on 3 frames and defined $K = 2$ to avoid discarding reviews, making it possible to follow the dataset predefined cross-validation splits (usual practice in experiments using this dataset (Araque et al., 2019)).

3.3.5. Results and Discussion

The average of the F1 score results of all the 10-fold runs of each trained model is shown in Table 3. We can figure out that the All Frames model obtains a better result than the Average Model, but not so expressively as in other scenarios. None of the Single Frame Models present a better F1 score than the Average Model. However, the Single Frame 2 Model, which represents the ending excerpt of the texts, is more successful in differentiating positive from negative reviews than the others. Analysing these findings, we can conclude that, even in this scenario that our approach does not achieve significantly better results than a summarised approach, it can suggest what portion of the text is more representative to a task.

[8]https://code.google.com/archive/p/word2vec/

3.3.6. Audio-Like Features Analysis

From the MCR boxplots of the positive flows shown in Figure 7, we can perceive that the negative reviews present higher MCR values than positive reviews in the first and second frames. In the third frame, the values are almost equal. The negative reviews instability regarding positive lexicon decreases from the first to the last frame, while the positive reviews show a peak in the second frame. If we consider the MCR boxplots of the negative flows (Figure 8), we realise that the situation is the opposite: the positive reviews instability is higher than the negative reviews and decreases over the frames.

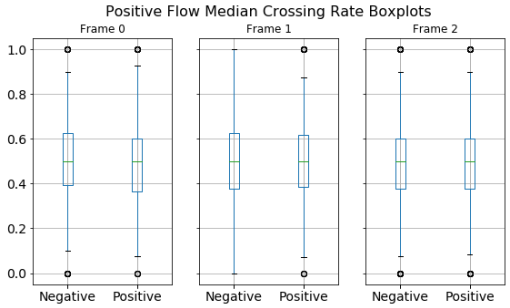

Figure 7: Boxplots of the Median Crossing Rate (MCR) for the Positive Flow.

4. Conclusions and Future Work

In this paper, we have introduced a promising novel approach to interpret human-generated texts representing and analysing them in their entirety, though not in a summarised way. More precisely, in order to represent texts, our model uses a sequence of information collected from the sentences of the text, which we called aspect flows. Then, inspired by audio analysis, this proposed model fragments the texts' aspect flows into frames and calculates Audio-Like Features for each one to perform text analysis. In the presented evaluation tasks, we have used aspect flows comprising the texts' sentences semantic distances to lexicons, considering

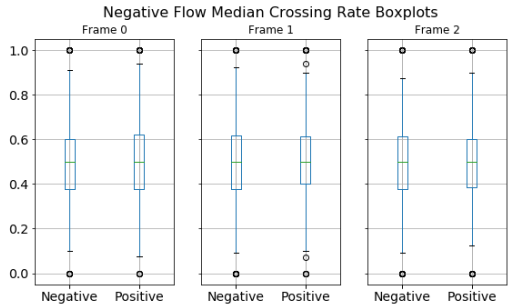

Figure 8: Boxplots of the Median Crossing Rate (MCR) for the Negative Flow.

a word embedding space. Using state-of-the-art machine learning classifiers, we have shown that this new approach outperforms the summarised features approach in different tasks that included diverse text kinds and domains and two distinct languages. Even when the results are not substantially better, our approach can evidence what portion of texts is more prone to differentiate them. Furthermore, we have shown that the investigation of the Audio-Like Features can reveal meaningful information about how each kind of text exploits an aspect, leading us to a deeper understanding of how these texts are written. As future work, we intend to implement frequency-domain features, after a criterion study about its viability, and mid-term features. We also plan to apply this method to other NLP tasks using larger texts.

5. Bibliographical References

Amorim, E., Cançado, M., and Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Araque, O., Corcuera-Platas, I., Sanchez-Rada, J. F., and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236 – 246.

Araque, O., Zhu, G., and Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346 – 359.

Bhattacharjee, M., Prasanna, S. R. M., and Guha, P. (2018). Time-frequency audio features for speech-music classification. *ArXiv*, abs/1811.01222.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

Filatova, E. (2017). Sarcasm detection using sentiment flow shifts. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, pages 264–269.

Giannakopoulos, T. and Pikrakis, A. (2014a). Chapter 1 - introduction. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 3 – 8. Academic Press, Oxford.

Giannakopoulos, T. and Pikrakis, A. (2014b). Chapter 2 - getting familiar with audio signals. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 9 – 31. Academic Press, Oxford.

Giannakopoulos, T. and Pikrakis, A. (2014c). Chapter 3 - signal transforms and filtering essentials. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 33 – 57. Academic Press, Oxford.

Giannakopoulos, T. and Pikrakis, A. (2014d). Chapter 4 - audio features. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 59 – 103. Academic Press, Oxford.

Helmholz, P., Meyer, M., and Robra-Bissantz, S. (2019). Feel the moosic: Emotion-based music selection and recommendation. 06.

Huang, G., Quo, C., Kusner, M. J., Sun, Y., Weinberger, K. Q., and Sha, F. (2016). Supervised word mover's distance. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4869–4877, USA. Curran Associates Inc.

Jalil, M., Butt, F., and Malik, A. (2013). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. pages 208–212, 05.

Jamdar, A., Abraham, J., Khanna, K., and Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence & Applications*, 6, 06.

Jeronimo, C., Marinho, L., Campelo, C., Veloso, A., and Melo, A. (2019). Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*.

Kaushik, L., Sangwan, A., and Hansen, J. (2013). Sentiment extraction from natural audio streams. 05.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lee, S.-W., Lee, J.-T., Song, Y.-I., and Rim, H.-C. (2010). High precision opinion retrieval using sentiment-relevance flows. pages 817–818, 01.

Lu, L. and Jiang, H. (2002). Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10:504 – 516, 11.

Mao, Y. and Lebanon, G. (2007). Isotonic conditional random fields and local sentiment flow. In B. Schölkopf, et al., editors, *Advances in Neural Information Processing Systems 19*, pages 961–968. MIT Press.

Nielsen, F. A. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA, October. Association for Computational Linguistics.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. In *PloS one*.

Seo, J. and Jeon, J. (2009). High precision retrieval using relevance-flow graph. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 694–695, New York, NY, USA. ACM.

Silverman, C. (2016). Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. buzzfeed, nov. 16.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Wachsmuth, H. and Stein, B. (2017). A universal model for discourse-level argumentation analysis. *ACM Trans. Internet Technol.*, 17(3):28:1–28:24, June.

Apêndice B

Classificação de Notícias Falsas em Português - Resultados de Todos os Cenários Executados

Neste Apêndice são apresentados os resultados de todos os cenários executados envolvendo a classificação de notícias falsas em português. A execução dos experimentos de todos os cenários seguem as configurações explanadas no Capítulo 5. O primeiro cenário foi discutido no Capítulo 5, mas os resultados são aqui repetidos para facilitar a comparação com os resultados dos demais.

B.0.1 Cenário Notícias Legítimas x Notícias Falsas

Neste cenário, o conjunto de dados de treino e de teste são constituídos por notícias legítimas de quaisquer domínios e fontes a que pertencem e todas as notícias falsas, que também são provenientes de uma variedade de fontes e domínios.

A Tabela B.1 apresenta as médias e desvios-padrão dos resultados das métricas Precision, Recall, F1-Score e PR-AUC das 500 rodadas de cada um dos modelos treinados obtidos para este cenário.

Tabela B.1: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação do cenário Notícias Legítimas x Notícias Falsas.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,29 \pm 0,13$	$0,31 \pm 0,17$	$0,40 \pm 0,15$	$0,34 \pm 0,14$	$0,43 \pm 0,13$	$0,44 \pm 0,13$
	XGB	$0,31 \pm 0,13$	$0,38 \pm 0,16$	$\mathbf{0,46 \pm 0,14}$	$0,38 \pm 0,12$	$0,47 \pm 0,11$	$\mathbf{0,49 \pm 0,13}$
Recall	RF	$0,08 \pm 0,04$	$0,06 \pm 0,03$	$0,11 \pm 0,05$	$0,10 \pm 0,05$	$0,14 \pm 0,05$	$0,14 \pm 0,06$
	XGB	$0,08 \pm 0,03$	$0,08 \pm 0,04$	$0,15 \pm 0,05$	$0,13 \pm 0,05$	$\mathbf{0,18 \pm 0,05}$	$0,17 \pm 0,06$
F1-Score	RF	$0,12 \pm 0,05$	$0,10 \pm 0,05$	$0,17 \pm 0,07$	$0,15 \pm 0,07$	$0,21 \pm 0,07$	$0,21 \pm 0,07$
	XGB	$0,12 \pm 0,05$	$0,12 \pm 0,05$	$0,23 \pm 0,07$	$0,19 \pm 0,06$	$0,26 \pm 0,07$	$0,25 \pm 0,07$
PR-AUC	RF	$0,26 \pm 0,03$	$0,24 \pm 0,03$	$0,33 \pm 0,04$	$0,29 \pm 0,04$	$0,33 \pm 0,04$	$0,33 \pm 0,05$
	XGB	$0,27 \pm 0,03$	$0,31 \pm 0,04$	$\mathbf{0,39 \pm 0,05}$	$0,33 \pm 0,04$	$0,39 \pm 0,05$	$\mathbf{0,40 \pm 0,05}$

B.0.2 Cenário *Cross-domain*

Neste cenário, são escolhidos domínios diferentes das notícias legítimas para formar os conjuntos de treino e teste no que diz respeito a essa classe de notícias. Por exemplo, o classificador pode ser treinado com notícias legítimas de Cultura e testado com notícias legítimas de Economia. Como há quatro domínios diferentes de notícias no *dataset*, os resultados

apresentados na Tabela B.2 se referem à média dos resultados obtidos nas doze combinações possíveis.

A Tabela B.2 apresenta as médias e desvios-padrão dos resultados das métricas Precision, Recall, F1-Score e PR-AUC das 500 rodadas de cada um dos modelos treinados obtidos para este cenário.

Tabela B.2: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação do cenário *Cross-domain*.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,26 \pm 0,11$	$0,26 \pm 0,14$	$0,34 \pm 0,14$	$0,34 \pm 0,17$	$0,45 \pm 0,15$	$0,39 \pm 0,16$
	XGB	$0,26 \pm 0,11$	$0,32 \pm 0,14$	$0,37 \pm 0,12$	$0,41 \pm 0,15$	$0,47 \pm 0,13$	$0,42 \pm 0,15$
Recall	RF	$0,10 \pm 0,05$	$0,08 \pm 0,04$	$0,13 \pm 0,06$	$0,08 \pm 0,04$	$0,12 \pm 0,05$	$0,10 \pm 0,05$
	XGB	$0,10 \pm 0,05$	$0,11 \pm 0,05$	$0,17 \pm 0,06$	$0,10 \pm 0,04$	$0,15 \pm 0,05$	$0,12 \pm 0,05$
F1-Score	RF	$0,14 \pm 0,06$	$0,11 \pm 0,06$	$0,18 \pm 0,08$	$0,12 \pm 0,06$	$0,19 \pm 0,07$	$0,15 \pm 0,07$
	XGB	$0,14 \pm 0,06$	$0,16 \pm 0,07$	$0,23 \pm 0,07$	$0,16 \pm 0,06$	$0,23 \pm 0,06$	$0,18 \pm 0,07$
PR-AUC	RF	$0,24 \pm 0,03$	$0,23 \pm 0,03$	$0,30 \pm 0,05$	$0,29 \pm 0,04$	$0,33 \pm 0,05$	$0,30 \pm 0,04$
	XGB	$0,26 \pm 0,04$	$0,28 \pm 0,05$	$0,35 \pm 0,06$	$0,34 \pm 0,04$	$0,38 \pm 0,05$	$0,35 \pm 0,05$

B.0.3 Cenário *Cross-source*

Neste cenário, o conjunto de treino é formado por notícias do Estadão e o de teste, por notícias da Folha de São Paulo, no tocante a notícias legítimas. O cenário foi executado em duas configurações: na primeira, não há distinção de domínio, logo todas as notícias são utilizadas para treino e teste. Na segunda, há distinção de domínio, ou seja, apenas um domínio é utilizado para classificação. Por exemplo, o treino utiliza as notícias do Estadão do domínio Cultura e o teste utiliza as notícias da Folha de São Paulo do mesmo domínio. Havendo quatro domínios diferentes de notícias no *dataset*, os resultados apresentados na Tabela B.4 se referem à média dos resultados obtidos nas quatro combinações possíveis.

A Tabela B.3 apresenta as médias e desvios-padrão dos resultados das métricas Precision, Recall, F1-Score e PR-AUC das 500 rodadas de cada um dos modelos treinados obtidos para este cenário na configuração sem distinção de domínio.

A Tabela B.4 apresenta as médias e desvios-padrão dos resultados das métricas Precision,

Tabela B.3: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação do cenário *Cross-source* sem distinção de domínio.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,32 \pm 0,14$	$0,31 \pm 0,17$	$0,43 \pm 0,16$	$0,36 \pm 0,16$	$0,45 \pm 0,15$	$0,44 \pm 0,14$
	XGB	$0,33 \pm 0,15$	$0,40 \pm 0,17$	$0,47 \pm 0,13$	$0,38 \pm 0,13$	$0,48 \pm 0,13$	$0,50 \pm 0,13$
Recall	RF	$0,07 \pm 0,03$	$0,06 \pm 0,03$	$0,11 \pm 0,05$	$0,10 \pm 0,05$	$0,14 \pm 0,05$	$0,14 \pm 0,06$
	XGB	$0,07 \pm 0,03$	$0,07 \pm 0,03$	$0,16 \pm 0,05$	$0,11 \pm 0,05$	$0,18 \pm 0,05$	$0,18 \pm 0,06$
F1-Score	RF	$0,11 \pm 0,05$	$0,09 \pm 0,05$	$0,17 \pm 0,07$	$0,15 \pm 0,07$	$0,21 \pm 0,07$	$0,21 \pm 0,07$
	XGB	$0,11 \pm 0,05$	$0,12 \pm 0,05$	$0,23 \pm 0,07$	$0,17 \pm 0,06$	$0,25 \pm 0,07$	$0,26 \pm 0,07$
PR-AUC	RF	$0,26 \pm 0,03$	$0,24 \pm 0,03$	$0,33 \pm 0,04$	$0,30 \pm 0,04$	$0,34 \pm 0,05$	$0,33 \pm 0,05$
	XGB	$0,28 \pm 0,03$	$0,31 \pm 0,04$	$0,40 \pm 0,05$	$0,33 \pm 0,04$	$0,40 \pm 0,05$	$0,41 \pm 0,05$

Recall, F1-Score e PR-AUC das 500 rodadas de cada um dos modelos treinados obtidos para este cenário na configuração com distinção de domínio.

Tabela B.4: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação do cenário *Cross-source* com distinção de domínio.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,35 \pm 0,15$	$0,33 \pm 0,17$	$0,43 \pm 0,19$	$0,29 \pm 0,18$	$0,40 \pm 0,18$	$0,31 \pm 0,17$
	XGB	$0,36 \pm 0,15$	$0,41 \pm 0,18$	$0,43 \pm 0,15$	$0,33 \pm 0,16$	$0,38 \pm 0,14$	$0,31 \pm 0,15$
Recall	RF	$0,10 \pm 0,05$	$0,07 \pm 0,04$	$0,14 \pm 0,08$	$0,07 \pm 0,05$	$0,11 \pm 0,06$	$0,09 \pm 0,05$
	XGB	$0,09 \pm 0,04$	$0,10 \pm 0,06$	$0,17 \pm 0,08$	$0,10 \pm 0,06$	$0,15 \pm 0,07$	$0,11 \pm 0,06$
F1-Score	RF	$0,15 \pm 0,06$	$0,11 \pm 0,06$	$0,20 \pm 0,10$	$0,11 \pm 0,07$	$0,17 \pm 0,08$	$0,13 \pm 0,07$
	XGB	$0,14 \pm 0,06$	$0,16 \pm 0,08$	$0,24 \pm 0,10$	$0,15 \pm 0,07$	$0,21 \pm 0,08$	$0,16 \pm 0,08$
PR-AUC	RF	$0,28 \pm 0,04$	$0,25 \pm 0,04$	$0,34 \pm 0,06$	$0,27 \pm 0,04$	$0,31 \pm 0,05$	$0,27 \pm 0,04$
	XGB	$0,30 \pm 0,04$	$0,32 \pm 0,06$	$0,39 \pm 0,07$	$0,31 \pm 0,05$	$0,36 \pm 0,06$	$0,30 \pm 0,05$

B.0.4 Cenário *Cross-source-domain*

Este cenário é uma mescla dos dois anteriores, ou seja, apenas notícias do Estadão são utilizadas para treino e da Folha de São Paulo para teste, porém são variados também os domínios de treino e teste. Por exemplo, o classificador pode ser treinado com notícias legítimas de

Cultura do Estadão e testado com notícias legítimas de Economia da Folha de São Paulo. Havendo quatro domínios diferentes de notícias no *dataset*, os resultados apresentados na Tabela B.5 se referem à média dos resultados obtidos nas doze combinações possíveis.

A Tabela B.5 apresenta as médias e desvios-padrão dos resultados das métricas Precision, Recall, F1-Score e PR-AUC das 500 rodadas de cada um dos modelos treinados obtidos para este cenário.

Tabela B.5: Média e desvio-padrão dos resultados das métricas **Precision, Recall, F1-Score** e **PR-AUC** da classificação do cenário *Cross-source-domain*.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Precision	RF	$0,28 \pm 0,13$	$0,29 \pm 0,15$	$0,38 \pm 0,16$	$0,29 \pm 0,17$	$0,41 \pm 0,15$	$0,32 \pm 0,15$
	XGB	$0,29 \pm 0,13$	$0,34 \pm 0,16$	$0,39 \pm 0,13$	$0,33 \pm 0,15$	$0,42 \pm 0,13$	$0,34 \pm 0,14$
Recall	RF	$0,10 \pm 0,04$	$0,07 \pm 0,04$	$0,14 \pm 0,08$	$0,07 \pm 0,04$	$0,12 \pm 0,05$	$0,09 \pm 0,05$
	XGB	$0,09 \pm 0,04$	$0,10 \pm 0,06$	$0,18 \pm 0,07$	$0,08 \pm 0,04$	$0,15 \pm 0,05$	$0,11 \pm 0,05$
F1-Score	RF	$0,13 \pm 0,06$	$0,11 \pm 0,06$	$0,20 \pm 0,10$	$0,11 \pm 0,06$	$0,18 \pm 0,07$	$0,14 \pm 0,07$
	XGB	$0,14 \pm 0,06$	$0,15 \pm 0,08$	$0,24 \pm 0,08$	$0,13 \pm 0,06$	$0,21 \pm 0,06$	$0,16 \pm 0,06$
PR-AUC	RF	$0,26 \pm 0,04$	$0,24 \pm 0,04$	$0,31 \pm 0,07$	$0,26 \pm 0,04$	$0,31 \pm 0,04$	$0,26 \pm 0,04$
	XGB	$0,30 \pm 0,04$	$0,29 \pm 0,06$	$0,36 \pm 0,07$	$0,31 \pm 0,04$	$0,37 \pm 0,05$	$0,30 \pm 0,05$

Apêndice C

Classificação de Sentimentos em Avaliações de Filmes - Resultados de Todas as Métricas

Neste Apêndice são apresentados os resultados de todas as métricas obtidas na execução dos experimentos envolvendo a classificação de sentimentos em avaliações de filmes. A execução de todos os experimentos seguem as configurações explanadas no Capítulo 5. A Tabela C.1 apresenta as médias e desvios-padrão dos resultados das métricas *Accuracy Precision, Recall, F1-Score* e *ROC-AUC* das 10 rodadas de cada um dos modelos treinados com fluxos obtidos a partir dos léxicos Afinn, BingLiu e SentiWordNet.

Tabela C.1: Média e desvio-padrão dos resultados das métricas **Accuracy, Precision, Recall, F1-Score** e **ROC-AUC** da classificação referente aos léxicos Afinn, BingLiu e SentiWordNet.

		Average Model (AVGM)	WMD Flow Model (WMDM)	All Frames Model (ALFM)	Single Frame Model 0 (SFM0)	Single Frame Model 1 (SFM1)	Single Frame Model 2 (SFM2)
Accuracy	Afinn	$0,70 \pm 0,03$	$0,67 \pm 0,03$	$0,72 \pm 0,01$	$0,62 \pm 0,03$	$0,63 \pm 0,04$	$0,70 \pm 0,03$
	BingLiu	$0,73 \pm 0,03$	$0,67 \pm 0,03$	$0,73 \pm 0,02$	$0,63 \pm 0,02$	$0,65 \pm 0,05$	$0,70 \pm 0,03$
	SentiWordNet	$0,63 \pm 0,04$	$0,58 \pm 0,04$	$0,67 \pm 0,04$	$0,57 \pm 0,03$	$0,58 \pm 0,03$	$0,66 \pm 0,03$
Precision	Afinn	$0,71 \pm 0,03$	$0,67 \pm 0,03$	$0,73 \pm 0,02$	$0,62 \pm 0,03$	$0,64 \pm 0,04$	$0,70 \pm 0,03$
	BingLiu	$0,74 \pm 0,03$	$0,68 \pm 0,03$	$0,74 \pm 0,03$	$0,64 \pm 0,02$	$0,66 \pm 0,05$	$0,71 \pm 0,03$
	SentiWordNet	$0,63 \pm 0,04$	$0,58 \pm 0,04$	$0,67 \pm 0,03$	$0,57 \pm 0,03$	$0,59 \pm 0,03$	$0,66 \pm 0,02$
Recall	Afinn	$0,69 \pm 0,04$	$0,66 \pm 0,02$	$0,71 \pm 0,03$	$0,61 \pm 0,05$	$0,62 \pm 0,04$	$0,68 \pm 0,05$
	BingLiu	$0,70 \pm 0,03$	$0,66 \pm 0,03$	$0,72 \pm 0,03$	$0,63 \pm 0,04$	$0,63 \pm 0,06$	$0,69 \pm 0,05$
	SentiWordNet	$0,63 \pm 0,05$	$0,57 \pm 0,06$	$0,65 \pm 0,05$	$0,56 \pm 0,04$	$0,57 \pm 0,04$	$0,67 \pm 0,06$
F1-Score	Afinn	$0,70 \pm 0,03$	$0,66 \pm 0,02$	$\mathbf{0,72 \pm 0,02}$	$0,62 \pm 0,04$	$0,63 \pm 0,04$	$0,69 \pm 0,03$
	BingLiu	$0,72 \pm 0,03$	$0,67 \pm 0,03$	$\mathbf{0,73 \pm 0,02}$	$0,63 \pm 0,03$	$0,64 \pm 0,05$	$0,70 \pm 0,03$
	SentiWordNet	$0,63 \pm 0,04$	$0,58 \pm 0,05$	$\mathbf{0,66 \pm 0,04}$	$0,56 \pm 0,04$	$0,58 \pm 0,03$	$0,66 \pm 0,04$
ROC-AUC	Afinn	$0,70 \pm 0,03$	$0,66 \pm 0,02$	$\mathbf{0,72 \pm 0,01}$	$0,62 \pm 0,03$	$0,63 \pm 0,04$	$\mathbf{0,70 \pm 0,03}$
	BingLiu	$0,73 \pm 0,03$	$0,67 \pm 0,03$	$0,73 \pm 0,02$	$0,63 \pm 0,02$	$0,65 \pm 0,05$	$0,70 \pm 0,03$
	SentiWordNet	$0,66 \pm 0,03$	$0,58 \pm 0,05$	$\mathbf{0,67 \pm 0,04}$	$0,57 \pm 0,03$	$0,58 \pm 0,03$	$0,66 \pm 0,03$