UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GIUSEPPE PORTOLESE

Uso de características textuais para a classificação multirrótulo de gêneros cinematográficos em Português

Maringá

2019

GIUSEPPE PORTOLESE

Uso de características textuais para a classificação multirrótulo de gêneros cinematográficos em Português

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.
Área de concentração: Ciência da Computação

Orientadora: Profª. Drª. Valéria Delisandra Feltrim

Maringá

2019

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Portolese, Giuseppe

P853u Uso de características textuais para a classificação multirrótulo de gêneros cinematográficos em Português / Giuseppe Portolese. -- Maringá, PR, 2019.
99 f.: il. color., figs., tabs.

Orientadora: Profa. Dra. Valéria Delisandra Feltrim.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2019.

1. Classificação multirrótulo. 2. Processamento da linguagem natural (Informática). 3. Base de dados (Informática). 4. Gêneros cinematográficos. I. Feltrim, Valéria Delisandra, orient. II. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 23.ed. 005.133

Elaine Cristina Soares Lira - CRB-9/1202

FOLHA DE APROVAÇÃO

GIUSEPPE PORTOLESE

Uso de características textuais para a classificação multirrótulo de gêneros cinematográficos em Português

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Profa. Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Marcos Aurélio Domingues
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Gustavo Henrique Paetzold
Universidade Tecnológica Federal do Paraná – COENC/UTFPR-TD

Aprovada em: 20 de setembro de 2019.
Local da defesa: Sala 101, Bloco C56, *campus* da Universidade Estadual de Maringá.

AGRADECIMENTOS

Agradeço primeiramente a meus pais Patrícia e Paulo por me proverem tudo que foi necessário para que eu chegasse até aqui. Todas as minhas realizações são mérito deles.

Agradeço à minha orientadora Valéria que me guiou em grande parte de minha jornada acadêmica, sendo uma das principais contribuidoras para que eu chegasse ao título de mestre.

Agradeço a meus professores por me darem a formação necessária, como acadêmico e como pessoa.

E meus amigos, que tornaram o peso dos problemas muito mais leve, e sem os quais eu não chegaria até aqui.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“The road to wisdom?
Well, it’s plain and simple to express:
Err and err and err again
but less and less and less.”

Piet Hein

Uso de características textuais para a classificação multirrótulo de gêneros cinematográficos em Português

RESUMO

Devido ao progresso tecnológico observado recentemente e a disponibilidade de grandes quantidades de dados, a necessidade da classificação automática de mídias digitais tem aumentado pois em muitos casos uma anotação manual é inviável devido ao tamanho de bases de tais dados. A classificação multirrótulo, na qual cada instância contida na base estudada pode ser classificada com diversos rótulos não-exclusivos, já é uma área estudada na literatura atual com diversos estudos avaliando problemas como a classificação de gêneros cinematográficos, no qual é comum que características audiovisuais sejam utilizadas para rotular filmes em gêneros pré-estabelecidos. No entanto, a análise de sinopses ainda é uma área pouco explorada nesse domínio, com o problema específico do estudo de sinopses na língua portuguesa recebendo pouca atenção. Neste trabalho são apresentadas três novas bases de dados de sinopses em português, sendo uma delas extraída de uma base internacional de filmes e as demais derivadas por rebalanceamento. Um total de 85 experimentos são conduzidos avaliando o uso de características provenientes de 9 grupos distintos em conjunto com 4 classificadores multirrótulo presentes na literatura, explorando técnicas de fusão tardia e imediata em experimentos individuais e combinatoriais. Valores de medida-F de $0,478$ para a base originalmente estudada, e $0,611$ para bases derivadas por rebalanceamento são obtidos, mostrando que os métodos utilizados são condizentes com aqueles encontrados no estado da arte da literatura.

Palavras-chave: Classificação Multirrótulo. Processamento de Linguagem Natural. Gêneros Cinematográficos.

On the usage of textual features for multi-label portuguese film genre classification

ABSTRACT

Due to the recent technological progress in recent years and to the availability of large data quantities, the need for the automatic classification of digital media has been increased as a result of the fact that a manual approach to such classification is unviable considering the size of such databases. Multi-label classification, in which each instance in a dataset can be classified as belonging to several non-exclusive labels is a domain that is already studied in the current literature, with several studies assessing the performance of tasks such as film genre classification, in which audiovisual features are commonly used to label films with a set of preestablished genres. Studies using approaches based on synopsis analysis are, however, much rarer in the current literature, with even fewer publications dedicated specifically to the study of Portuguese language synopses. In this work we present three new Portuguese language synopses datasets, with one of them extracted from an international movie database and the remaining two being resampled versions of the original one. A total of 85 experiments were conducted, assessing the performance of features from 9 different groups when used in combination with 4 multi-label classifiers, exploring early and late fusion techniques in individual and combinatorial experiments. Results from the conducted experiments present F-measure scores of up to $0,478$ for the original dataset and $0,611$ for the resampled datasets, demonstrating that the implemented methods have similar performance to those found in the state of the art of the movie genre classification task.

Keywords: Multi-label Classification. Natural Language Processing. Film Genres.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1

INTRODUÇÃO

Segundo Herrera et al. (2016), o progresso tecnológico observado nos últimos anos proporcionou um crescimento exponencial em termos de armazenamento e distribuição de dados, aumentando a necessidade de processamento automático de tais informações. Os autores destacam que a classificação automática de diversos tipos de informação digital como textos, fotos, música e vídeos tem uma demanda crescente e que o domínio da classificação multirrótulo prove métodos que permitem tal classificação em diversas categorias não-exclusivas.

Dentre os possíveis tipos de mídia digital que podem ser classificados com diversos gêneros não-exclusivos estão obras cinematográficas. Estudos sobre a classificação de gêneros de filmes já são presentes na literatura, sendo comum abordarem o problema por meio do uso de características extraídas de seu conteúdo visual, como em Rasheed et al. (2005), Zhou et al. (2010) e Wehrmann e Barros (2017). Também há estudos que utilizam características extraídas de sinopses de filmes, como em Hoang (2018) e Ho (2011), que utilizam sinopses escritas em língua inglesa e Rahman et al. (2017), que abordou a classificação de filmes de origem indiana.

Apesar de haverem estudos que utilizam características textuais extraídas das sinopses, não foram encontrados trabalhos que aprofundassem a avaliação quanto à contribuição de características provenientes de diversos tipos distintos de extração, nem de suas possíveis combinações. Tampouco se encontrou trabalhos que abordassem a classificação de gêneros de filmes utilizando sinopses escritas em língua portuguesa. Dessa forma, este trabalho abordou esse problema ao explorar combinações entre os diversos métodos de classificação presentes na literatura, usando-se de diversos conjuntos de características e classificadores para aferir a qualidade de sua classificação no subdomínio da classificação multirrótulo de gêneros

cinematográficos utilizando características textuais extraídas de sinopses. Uma nova base de dados composta de sinopses em português é apresentada em conjunto com duas de suas versões alternativas derivadas por rebalanceamento. Os resultados de classificação foram comparados com o estado da arte da classificação de gêneros cinematográficos usando características de suas sinopses para que particularidades sobre as bases de dados, conjuntos de características e classificadores utilizados pudessem ser avaliadas, visando um entendimento maior sobre o domínio como um todo.

Resultados obtidos neste trabalho mostraram-se condizentes com o estado da arte do domínio, sendo que em casos nos quais algoritmos de rebalanceamento da base original foram utilizados, os métodos de classificação utilizados puderam superar todos os apresentados nos trabalhos diretamente relacionados a este. Experimentos com diversos classificadores e conjuntos de características diferentes foram produzidos, mostrando que métodos simples como a classificação baseada em Relevância Binária e conjuntos de características baseados na medida TF-IDF (*Term Frequency-Inverse Document Frequency*) apresentaram-se como os métodos mais adequados para a realização de classificações no domínio atual, mesmo que métodos mais sofisticados também tenham sido estudados.

O presente trabalho é então organizado da seguinte forma: Na Seção 2 é apresentada a revisão bibliográfica, formando a base teórica requerida para a realização deste estudo. Em seguida na Seção 3 são explorados trabalhos relacionados que abordam o problema da classificação de gêneros cinematográficos, sejam eles direta ou indiretamente relacionados ao presente estudo. Na Seção 4 é discutido o desenvolvimento do trabalho, explorando a base de dados P-TMDb, que foi extraída e preprocessada para uso deste estudo, bem como as estratégias de rebalanceamento exploradas e que geraram as bases P-TMDb(-) e P-TMDb(+). Um experimento realizado com anotadores humanos utilizando um subconjunto da base original também é discutido e são detalhados os experimentos realizados com os algoritmos classificadores. Na Seção 5 os resultados de todos os experimentos realizados neste estudo são apresentados e discutidos. E, por fim, na Seção 6 são apresentadas as conclusões deste estudo, assim como direcionamentos para possíveis trabalhos futuros.

2

FUNDAMENTAÇÃO TEÓRICA

Nesta seção são explorados diversos trabalhos publicados atualmente na literatura, afim de apresentar uma base teórica para o melhor entendimento do estudo apresentado neste documento. Tópicos gerais sobre aprendizagem de máquina são introduzidos e, uma vez que uma base teórica suficiente seja apresentada, definições formais acerca dos tópicos abordados neste trabalho são introduzidas. Em seguida algoritmos e modelos empregados ao longo deste documento são apresentados, em conjunto com suas definições, equações, e pseudocódigos quando necessário.

2.1 MINERAÇÃO DE DADOS

Segundo Herrera et al. (2016) devido à vasta disponibilidade de informações de milhões de usuários e suas interações com instituições e companhias, a tarefa de extração de informações úteis acerca de tais interações de forma manual tornou-se extremamente difícil no contexto atual. Devido a tal fato técnicas de mineração de dados, também conhecida como *Data Mining* (DM), tem tornado-se populares por proverem formas automáticas de extração de conhecimento em bases de dados que possam auxiliar sistemas de tomada de decisões, descreverem estruturas de informação ou prever em dados futuros.

Os autores citam ainda que a mineração de dados é uma das etapas de um processo conhecido como descoberta de conhecimento em bancos de dados, também chamada de *Knowledge Discovery in Databases* (KDD), que é descrito como um processo não-trivial de identificação de padrões válidos, úteis e aproveitáveis em grandes quantidades de dados. A

descoberta de conhecimento em bancos de dados é composta de em diversas etapas, como apresentado na Figura 2.1. As etapas do processo são:

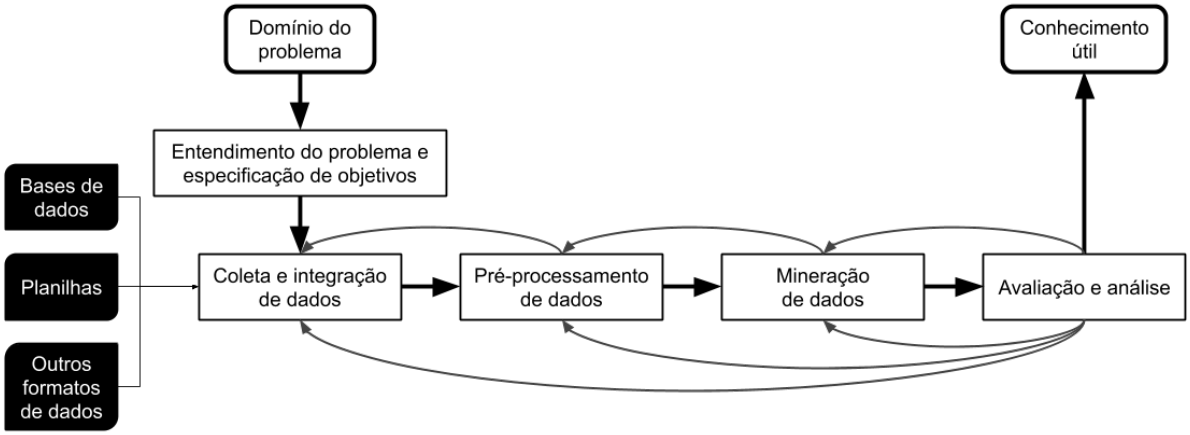
1. **Coleta e integração de dados:** Representa a coleta de dados de uma ou mais fontes como bases de dados, planilhas e outros arquivos de dados. Uma vez que tais informações podem ser representados de maneiras distintas, nesta etapa todos os dados são integrados em uma representação comum, adequada para a realização das etapas seguintes.

2. **Pré-processamento de dados:** Devido ao fato que algumas das informações coletadas possam ser inconsistentes, irrelevantes, possam apresentar diferenças de escala, ruídos ou outras anomalias, esta etapa consiste na limpeza dos dados para que apenas informações verdadeiramente relevantes possam ser utilizadas por etapas seguintes.

3. **Mineração de dados:** Descrita por Herrera et al. (2016) como etapa ocasionalmente considerada como principal da descoberta de conhecimento em bancos de dados, a mineração de dados tem como objetivo a execução de algoritmos que possam extrair conhecimento dos dados provenientes das etapas anteriores. Algoritmos utilizados nesta etapa podem realizar tarefas como agrupamento de dados baseando-se em seus atributos ou produzir modelos capazes de classificar automaticamente dados futuros, entre outras tarefas que serão descritas à seguir.

4. **Avaliação e análise:** Consiste da interpretação de resultados obtidos da etapa anterior, com o objetivo de ajudar o usuário a realizar os objetivos relacionados ao domínio do problema e gerando conhecimento útil acerca do mesmo.

Figura 2.1: Etapas do processo de descoberta de conhecimento em bancos de dados



Fonte: Adaptado de Herrera et al. (2016)

Herrera et al. (2016) notam que, segundo o que pode ser observado no diagrama da Figura 2.1, o processo descrito pode retroceder em suas etapas de acordo com condições encontradas durante sua realização, sendo que a iteração de cada etapa visa melhorar a qualidade do conhecimento gerado ao final do processo.

Ainda segundo Herrera et al. (2016), algoritmos utilizados na etapa de mineração de dados podem ser agrupados segundo diversos critérios, sendo que entre eles estão o tipo de rotulação presente nos dados estudados, tipo de resultado esperado do processo, e modelo utilizado para a representação do conhecimento.

Os autores notam que a natureza dos dados pode afetar os métodos de mineração de dados que podem ser aplicados a eles, como também determinar quais objetivos podem ser definidos para o processo de aprendizagem. A Figura 2.2 representa exemplos das três principais categorias de métodos quanto à natureza dos rótulos atribuídos a cada uma das instâncias estudadas. As categorias exemplificadas pela figura são:

- **Aprendizado supervisionado:** Cada instância da base de dados foi previamente rotulada por especialistas no domínio. Métodos de mineração podem então usar esse tipo de informação para atribuir rótulos a novas instâncias de acordo com seus atributos, sendo que neste contexto um rótulo pode ser um valor numérico contínuo ou discreto atribuído a cada instância na base de dados.

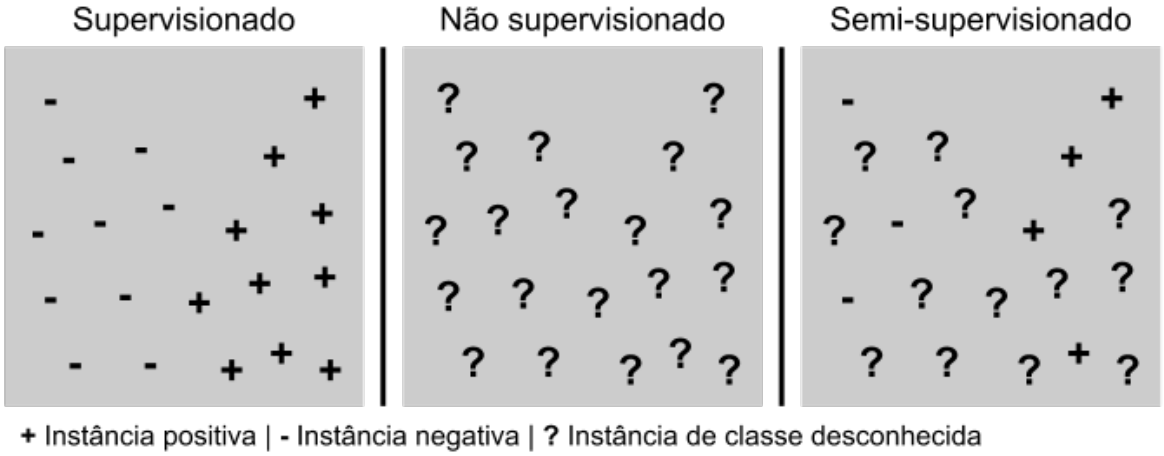
- **Aprendizado não supervisionado:** Nenhuma das instâncias de dados estão rotuladas, sendo representadas apenas por seus atributos. Métodos que trabalham com dados desse tipo focam em agrupar instâncias usando-se dos valores de seus atributos, ou em identificar padrões que possam estar presentes nos dados.

- **Aprendizado semi-supervisionado:** Instâncias da base de dados podem estar rotuladas ou não possuírem rótulo. Métodos de mineração podem combinar técnicas das duas categorias anteriores, prevendo modelos de dados que podem ser aprimorados com informações de instâncias não rotuladas, ou analogamente aprimorando agrupamentos não supervisionados ao introduzir informações obtidas de instâncias rotuladas.

Herrera et al. (2016) também citam que métodos de mineração podem diferir quanto aos objetivos buscados, sendo que tais objetivos podem ser classificados em duas categorias principais:

- **Tarefas preditivas:** O objetivo deste tipo de tarefa é inferir um valor de saída a partir dos atributos fornecidos como entradas. Esse tipo de tarefa requer que instâncias de entradas

Figura 2.2: Principais tipos de aprendizado de acordo com a natureza de seus dados



Fonte: Adaptado de Herrera et al. (2016)

sejam rotuladas, ligando a tarefa ao aprendizado supervisionado e semi-supervisionado. Algumas das principais tarefas preditivas são: classificação, na qual rótulos categóricos são atribuídos a novas instâncias; e regressão, na qual valores reais são estimados para as novas instâncias.

- **Tarefas descritivas:** Tarefas nas quais o objetivo é a extração de informações acerca de informações estruturais da base estudada. Tarefas desse tipo são comumente ligadas ao aprendizado não-supervisionado, porém podem ainda ser aplicadas ao aprendizado supervisionado e semi-supervisionado. Algumas das principais tarefas descritivas são: agrupamento de instâncias; geração de regras de associação; detecção de anomalias nos dados; e redução de dimensionalidade.

Herrera et al. (2016) notam que a tarefa de classificação está entre as mais populares na literatura, sendo que nesse tipo de tarefa os algoritmos fazem uso de dois conjuntos de atributos em seu funcionamento: atributos de entrada, compostos por variáveis descrevendo cada instância da base de dados; e atributos de saída, representando rótulos ou classes atribuídas a cada instância da base. A Tabela 2.1 apresenta as diversas classificações de tarefa de classificação descritas a seguir:

- **Classificação binária:** Citada por Herrera et al. (2016) como o problema mais simples de classificação, consiste em atribuir um valor binário a cada instância classificada. Exemplos de aplicações que fazem uso de classificação binária são: filtragem de *e-mails*; análise de crédito; avaliação médica; e reconhecimento de diversos tipos de padrões binários.

- **Classificação multiclasse:** Nesse tipo de classificação, a saída única atribuída a cada nova instância pode pertencer a um conjunto de valores possíveis, sendo que o significado e tipo de tais valores é específico para cada aplicação. A classificação multiclasse pode ser vista como uma generalização da classificação binária, pois esta pode ser considerada como uma classificação multiclasse na qual apenas dois valores distintos pode ser usado na classificação de cada instância. Um dos exemplos para classificação multiclasse é o da identificação de espécies de flores.

- **Classificação multirrótulo:** Diferencia-se das classificações anteriores por atribuir a cada nova instância um vetor de rótulos de tamanho fixo, sendo que o tamanho de tal vetor é determinado pelo número de classes presentes na base de dados estudada. Cada elemento do vetor de saída é um valor binário indicando a relevância de cada rótulo para a instância classificada. Dessa forma, cada instância pode ter diversos rótulos atribuídos a ela ao mesmo tempo, sendo que cada combinação única possível de rótulos ativos e inativos em uma instância é chamado *labelset* (conjunto de rótulos). A classificação multirrótulo é comumente utilizada em problemas de classificação de imagens, música e vídeo.

- **Classificação multidimensional:** Pode ser vista como uma generalização da classificação multirrótulo, tendo um vetor de saída para cada instância, porém com cada elemento de tal vetor podendo conter valores de um conjunto pré-determinado de saídas. Tal classificação pode ser utilizada em domínios similares ao da classificação multirrótulo, prevendo uma quantidade maior de valores de saída para cada rótulo na base.

- **Classificação multi-instância:** Difere de todas as outras classificações por atribuir valores de classe para grupos de instâncias de entrada. Os autores citam que um dos tipos de aplicação mais facilmente interpretáveis de classificação multi-instância é a de categorização de imagens, na qual cada imagem é separada em regiões menores que são analisadas como instâncias distintas e então utilizadas para classificar a imagem como um todo.

2.2 CLASSIFICAÇÃO MULTIRRÓTULO

Segundo Herrera et al. (2016), a classificação multirrótulo é uma tarefa preditiva da mineração de dados na qual os modelos construídos devem classificar entradas que podem possuir vários

Tabela 2.1: Problemas de classificação de acordo com saídas previstas

Quantidade de saídas	Tipo das saídas	Tipo da Classificação
1 por instância	Binárias	Binária
1 por instância	Multivaloradas	Multiclasse
n por instância	Binárias	Multirrótulo
n por instância	Multivaloradas	Multidimensional
1 por n instâncias	Binárias/Multivaloradas	Multi-instância

Fonte: Adaptado de Herrera et al. (2016)

rótulos ao mesmo tempo. Os autores destacam que a classificação multirrótulo é utilizada em diversas aplicações do mundo real, tais como a classificação automática de textos, imagens, música e vídeo, que comumente possuem diversos rótulos com os quais são classificados ao mesmo tempo. Tais rótulos podem representar informações sobre itens a que estão atribuídos dependendo do domínio estudado, como gêneros cinematográficos ou musicais, tópicos de notícias, e tipos de paisagem presentes em uma fotografia.

Para definir formalmente um classificador multirrótulo ($\mathcal{F}$), Herrera et al. (2016) apresentam as definições formais de espaço de entrada ($\mathcal{X}$), conjunto de rótulos ($\mathcal{L}$), espaço de saída ($\mathcal{Y}$) e base de dados multirrótulo ($\mathcal{D}$), conforme segue.

1. Seja $\mathcal{X}$ o espaço de entrada da classificação, com amostras $X \in A_1 \times A_2 \times ... \times A_f$, sendo f o número de atributos da entrada e $A_1, A_2, ..., A_f$ conjuntos arbitrários. Portanto, cada instância X é dada pelo produto cartesiano entre esses conjuntos.

2. Seja $\mathcal{L}$ o conjunto de todos os rótulos possíveis. $\mathcal{P}(\mathcal{L})$ denota o conjunto potência (*powerset*) de $\mathcal{L}$, que contém todas as possíveis combinações de rótulos $l \in \mathcal{L}$, incluindo o conjunto vazio e $\mathcal{L}$. $k = |\mathcal{L}|$ é o número total de rótulos em $\mathcal{L}$.

3. Seja $\mathcal{Y}$ o espaço de saída da classificação, contendo todos os possíveis vetores (*labelsets*) Y, $Y \in \mathcal{P}(\mathcal{L})$. O tamanho de Y sempre será k.

4. Seja $\mathcal{D}$ uma base de dados multirrótulo, contendo um subconjunto finito de $A_1 \times A_2 \times ... \times A_f \times \mathcal{P}(\mathcal{L})$. Cada elemento $(X,Y) \in \mathcal{D} | X \in A_1 \times A_2 \times ... \times A_f, Y \in \mathcal{P}(\mathcal{L})$ será uma instância ou entrada da base. $n = |\mathcal{D}|$ é o número de elementos em $\mathcal{D}$.

5. Seja $\mathcal{F}$ um classificador multirrótulo definido como $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$. A entrada de $\mathcal{F}$ será uma instância qualquer $X \in \mathcal{X}$, e sua saída será uma previsão $Z \in \mathcal{Y}$. Portanto, o vetor de previsão contendo rótulos para uma instância qualquer pode ser obtido por $Z = \mathcal{F}(X)$.

Uma vez que na classificação multirrótulo a saída prevista é um vetor de rótulos e não um rótulo único, como na classificação tradicional (binária e multiclasse), é necessária a utilização de métricas próprias para esse domínio, seja para caracterizar a base $\mathcal{D}$ ou para avaliar o classificador $\mathcal{F}$. Dessa forma, podemos então construir uma lista de símbolos descritos por Herrera et al. (2016) que podem ser utilizados para a apresentação da métricas específicas para o contexto da classificação multirrótulo apresentadas a seguir, sendo que tais símbolos são:

- $\mathcal{D}$ Uma base de dados multirrótulo qualquer;
- n Número de entradas em $\mathcal{D}$;
- $\mathcal{L}$ Conjunto completo de rótulos presentes em $\mathcal{D}$;
- l Qualquer rótulo de $\mathcal{L}$;
- k Número total de elementos em $\mathcal{L}$;
- X Conjunto de atributos de entrada para uma instância qualquer;
- f Número de atributos contidos em X;
- $\mathcal{X}$ Espaço de entrada de $\mathcal{D}$, formado por todas as entradas X;
- Y Conjunto de rótulos de saída (*labelset*) de uma entrada qualquer;
- $\mathcal{Y}$ Espaço de saída de $\mathcal{D}$, formado por todas as saídas Y;
- Z O conjunto de rótulos de saída previsto pelo classificador;

A primeira métrica, chamada cardinalidade, diz respeito a bases de dados multirrótulos e foi inicialmente apresentada por Tsoumakas e Katakis (2007). A cardinalidade de uma base D é dada pela Equação 1 e representa o número médio de rótulos atribuídos as instâncias da base em questão. Valores de cardinalidade próximos a 1 denotam que, em geral, cada instância é classificada com poucos rótulos, enquanto que valores próximos a k denotam que a base estudada é verdadeiramente multirrótulo e possui diversos rótulos atribuídos a cada instância.

$$Card(D) = \frac{1}{n}\sum_{i=1}^{n}|Y_i| \tag{1}$$

A segunda métrica, chamada densidade, também diz respeito à base de dados e é derivada da cardinalidade. A densidade de uma base D é dada pela Equação 2 e corresponde à cardinalidade normalizada pelo número de rótulos presentes na base de dados. Enquanto a

cardinalidade provê o número médio de rótulos ativos por entrada, a densidade provê o número relativo de rótulos ativos, representando mais claramente a medida em que a base de dados possui rótulos atribuídos a cada entrada.

$$Dens(D) = \frac{1}{k}\frac{1}{n}\sum_{i=1}^{n}|Y_i| \tag{2}$$

Outra métrica utilizada para medir o quanto uma base de dados é multirrótulo é a medida de Percentual Mínimo (P_{min}) introduzida por Turner et al. (2013). A medida P_{min} é dada pela Equação 3 e representa a proporção de entradas da base que possuem apenas um rótulo. Dessa forma, um valor alto de P_{min} implicaria em uma base de dados majoritariamente monorrótulo.

$$P_{min}(D) = \sum_{y' \in Y / |y'|=1} \frac{|y'|}{n} \tag{3}$$

As demais métricas referentes às bases de dados multirrótulo utilizadas neste estudo foram inicialmente propostas por Charte et al. (2015) e dizem respeito ao desbalanceamento da distribuição de rótulos na base.

A métrica *IRLbl* (*Imbalance Ratio per Label*) avalia o desbalanceamento de um rótulo l por meio da proporção entre a frequência do rótulo mais comum e a frequência de l. A *IRLbl* do rótulo mais frequente será 1, enquanto que a *IRLbl* de todos os outros rótulos será maior que 1. Dessa forma, quanto maior for o valor *IRLbl* de um rótulo, mais raro esse rótulo é na base de dados. A medida *IRLbl* é dada pela Equação 4, na qual os operadores $[[expr]]$ representam colchetes de Iverson, que retornam 1 se *expr* for verdadeira ou 0 caso contrário.

$$IRLbl(l) = \frac{\max_{l' \in L}(\sum_{i=1}^{n}[[l' \in Y_i]])}{\sum_{i=1}^{n}[[l \in Y_i]]} \tag{4}$$

Após o cálculo do *IRLbl* de cada rótulo, outras métricas podem ser obtidas para que se examine o grau de desbalanceamento da base de dados. A métrica *MaxIR*, definida pela Equação 5, representa a maior proporção de desbalanceamento da base, mostrando a proporção da frequência do rótulo mais raro em relação ao mais comum. Já a métrica *MeanIR*, definida pela Equação 6, calcula a proporção de desbalanceamento média na base, que denota, de uma maneira geral, o grau de desbalanceamento médio entre todos os rótulos.

$$MaxIR = \max_{l \in L}(IRLbl(l)) \tag{5}$$

$$MeanIR = \frac{1}{k} \sum_{l \in L} IRLbl(l) \tag{6}$$

Cabe destacar que, enquanto a métrica *MeanIR* pode dar a medida geral de desbalanceamento da base, seus valores podem ser afetados tanto por altos valores de *IRLbl* para vários rótulos, quanto por valores extremos de *IRLbl* para poucos rótulos. Tendo isso em mente, Charte et al. (2015) apresentam também a métrica de dispersão *CVIR*, representada nas Equações 7 e 8, que visa calcular o coeficiente de variação entre as proporções de desbalanceamento por rótulo. Os autores notam que, quanto maior o *CVIR* de uma base, maior é a variação entre as proporções de desbalanceamento, indicando que a *MeanIR* é afetada majoritariamente por valores extremos de *IRLbl* para poucos rótulos.

$$IRLbl\sigma = \sqrt{\frac{1}{k-1} \sum_{l \in L} (IRLbl(l) - MeanIR)^2} \tag{7}$$

$$CVIR = \frac{IRLbl\sigma}{MeanIR} \tag{8}$$

Por fim, neste estudo também foi feito o uso de métricas específicas para aferir o desempenho dos classificadores. As métricas apresentadas a seguir são calculadas para cada instância i das n instâncias de teste e comparam os conjuntos Y_i e Z_i, que representam, respectivamente, o conjunto de rótulos associado à entrada i da base e o conjunto de rótulos atribuídos à entrada i pelo classificador.

A métrica de precisão para uma classificação multirrótulo é dada pela Equação 9 e representa a média das proporções entre o número de rótulos corretamente atribuídos às instâncias em relação ao total de rótulos atribuídos pelo classificador.

$$Precisão = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{Z_i} \tag{9}$$

A métrica revocação é dada pela Equação 10 e representa a média das proporções entre o número de rótulos corretamente atribuídos às instâncias em relação ao total de rótulos das instâncias.

$$Revocão = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{Y_i} \tag{10}$$

A medida-F é uma forma de combinar os valores de precisão e revocação como um

métrica única. Dada pela Equação 11, a métrica corresponde à média harmônica entre a precisão e a revocação.

$$Medida - F = 2 * \frac{Precisão * Revocão}{Precisão + Revocão} \tag{11}$$

Os valores de precisão, revocação e medida-F calculados pelas equações acima são obtidos por instância, sendo o valor médio obtido pela divisão pelo número de instâncias (n). Essas mesmas métricas podem ser calculadas por rótulo, considerando-se, para um rótulo l, os valores de verdadeiros positivos (TP_l), falsos positivos (FP_l), verdadeiros negativos (TN_l) e falsos negativos (FN_l). Segundo Herrera et al. (2016), essas métricas podem ter duas variações distintas dependendo da forma como é calculado o respectivo valor médio. Versões *macro* das métricas podem ser construídas de forma que todos os rótulos possuam o mesmo peso no cálculo do valor da métrica para a base como um todo, enquanto que versões *micro* visam combinar as métricas de cada rótulo de maneira que a distribuição dos rótulos na base afete os pesos utilizados na combinação. Neste trabalho, a versão *micro* das métricas precisão, revocação e medida-F foi considerada.

2.3 REAMOSTRAGEM DE BASES MULTIRRÓTULO

Segundo Charte et al. (2015), o problema de desbalanceamento em bases multirrótulo se dá quando há diferenças nas frequências dos rótulos da base. Os autores citam que o problema é profundamente estudado na literatura, uma vez que é comum que bases multirrótulo apresentem altos níveis de desbalanceamento. Charte et al. (2015) então propõem alguns algoritmos de pré-processamento que podem beneficiar a classificação de bases multirrótulo por modificarem suas entradas de modo a tornar a base, como um todo, mais balanceada.

Métodos de reamostragem de bases de dados podem ser divididos em métodos de *oversampling*, que adicionam instâncias à base, e de *undersampling*, que removem instâncias da base. Eles podem ainda ser classificados como métodos aleatórios, que fazem a adição/remoção de instâncias pode ser feita de forma aleatória, ou métodos heurísticos, que usam alguma(s) heurística(s) no processo de reamostragem. Neste trabalho optou-se pelo uso dos algoritmos de reamostragem aleatória propostos por Charte et al. (2013) chamados LP-RUS (*Label Powerset Random Undersampling)* e LP-ROS (*Label Powerset Random Oversampling*). Ambos algoritmos fazem a reamostragem analisando os conjuntos de rótulos (*labelsets*) presentes na base.

O algoritmo LP-RUS, descrito no Algoritmo 1, faz a reamostragem da base original por meio da remoção de instâncias aleatórias cujos *labelsets* sejam majoritários na base. O processo é então repetido até que o número de entradas na nova base derivada seja reduzido por um determinado valor ou que todos os *labelsets* da base tenham igual representação.

Algoritmo 1: Pseudocódigo do algoritmo LP-RUS

Entrada: Base de dados D, Porcentagem P

Saída: Base de dados pré-processada

$entradasADeletar \leftarrow |D|/100 * P$; `// Redução de P% do tamanho`

`/* Agrupando entradas de acordo com seus labelsets */`

para $i = 1 \rightarrow |labelsets|$ **faça**
  $bolsaLabelSet_i \leftarrow entradasComLabelset(i)$
fim

`/* Calculando a média de entradas por labelset */`

$tamanhoMédio \leftarrow 1/|labelsets| * \sum_{i=1}^{|labelsets|} |bolsaLabelSet_i|$

`/* Obtendo as Bolsas de LabelSets majoritárias */`

para cada $bolsaLabelSet_i \in bolsaLabelSet$ **faça**
  se $|bolsaLabelSet_i| > tamanhoMédio$ **então**
    $bolsaMajoritária_i \leftarrow bolsaLabelSet_i$
  fim
fim

$reduçãoMédia \leftarrow entradasADeletar/|bolsaMajoritária|$

$bolsaMajoritária \leftarrow OrdenaDoMenorParaOMaior(bolsaMajoritária)$

`/* Calculando número de instâncias a remover e removendo-as */`

para cada $bolsaMajoritária_i \in bolsaMajoritária$ **faça**
  $rBolsa_i \leftarrow min(|bolsaMajoritária_i| - tamanhoMédio, reduçãoMédia)$
  $restante \leftarrow reduçãoMédia - rBolsa_i$
  $distribuirEntreBolsas_{j>i}(restante)$
  para $n = 1 \rightarrow rBolsa_i$ **faça**
    $x \leftarrow random(1, |majBolsa_i|)\ deletarEntrada(bolsaMajoritária_i, x)$
  fim
fim

Fonte: Adaptado de Charte et al. (2015)

Charte et al. (2015) notam que o algoritmo LP-RUS, apesar de levar em consideração o grau de desbalanceamento de todos os rótulos, remove entradas pertencentes a diversas combinações de rótulos e não somente aos rótulos mais desbalanceadas.

De forma semelhante ao algoritmo LP-RUS, o algoritmo LP-ROS duplica instâncias aleatórias dos *labelsets* menos frequentes de forma a aumentar o tamanho da base original D até que essa seja incrementada em $P\%$. Os autores descrevem que o procedimento de *oversampling* do algoritmo ocorre de forma análoga ao funcionamento do algoritmo LP-RUS, mas que uma coleção de grupos da minoria $bolsaMinoritária_i$ com $|bolsaLabelSet_i| < tamanhoMédio$ é obtido, um $incrementoMédio = entradasADuplicar/bolsaMinoritária$ é calculado, e ao processar os grupos de minoria do maior para o menor um incremento individual $bolsaMinoritária_i$ é determinado. Os autores notam que, se $bolsaMinoritária_i$ alcança $tamanhoMédio$ antes que o número requerido de instâncias tenham sido duplicadas, o excesso é então distribuído entre as $bolsaMinoritária$ restantes, fazendo com que *labelsets* com menor representação tenham um maior número de clones e ajustando a sua frequência para chegar a uma base mais uniforme. Uma vez que tais considerações sejam feitas, um possível pseudocódigo para o algoritmo LP-ROS pode ser construído como exemplificado no Algoritmo 2.

Algoritmo 2: Pseudocódigo do algoritmo LP-ROS

Entrada: Base de dados D, Porcentagem P

Saída: Base de dados pré-processada

$entradasADuplicar \leftarrow |D|/100 * P$; // Incremento de $P\%$ do tamanho

/* Agrupando entradas de acordo com seus *labelsets* */

para $i = 1 \rightarrow |labelsets|$ **faça**
  $bolsaLabelSet_i \leftarrow entradasComLabelset(i)$
fim

/* Calculando a média de entradas por *labelset* */

$tamanhoMédio \leftarrow 1/|labelsets| * \sum_{i=1}^{|labelsets|} |bolsaLabelSet_i|$

/* Obtendo as Bolsas de *LabelSets* minoritárias */

para cada $bolsaLabelSet_i \in bolsaLabelSet$ **faça**
  se $|bolsaLabelSet_i| < tamanhoMédio$ **então**
    $bolsaMinoritária_i \leftarrow bolsaLabelSet_i$
  fim
fim

$incrementoMédio \leftarrow entradasADuplicar/|bolsaMinoritária|$

$bolsaMinoritária \leftarrow OrdenaDoMaiorParaOMenor(bolsaMinoritária)$

/* Calculando número de instâncias a duplicar e duplicando-as */

para cada $bolsaMinoritária_i \in bolsaMinoritária$ **faça**
  $iBolsa_i \leftarrow min(|bolsaMinoritária_i| - tamanhoMédio, incrementoMédio)$
  $restante \leftarrow incrementoMédio - iBolsa_i$
  $distribuirEntreBolsas_{j>i}(restante)$
  para $n = 1 \rightarrow iBolsa_i$ **faça**
    $x \leftarrow random(1, |minBolsa_i|)$ $duplicarEntrada(bolsaMinoritária_i, x)$
  fim
fim

Fonte: Adaptado de Charte et al. (2015)

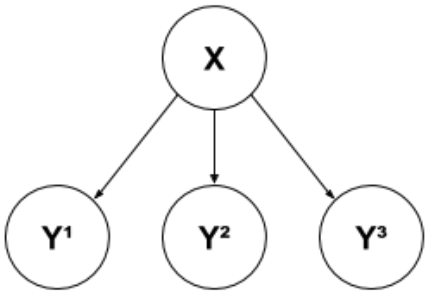
2.4 ALGORITMOS DE CLASSIFICAÇÃO MULTIRRÓTULO

Nesta subseção são apresentados os algoritmos de classificação multirrótulo presentes na literatura utilizados neste trabalho. Detalhes sobre o funcionamento de cada algoritmo são dados para que análises sobre sua performance possam ser observados nas seções seguintes.

2.4.1 Relevância Binária

Segundo Read et al. (2011), o algoritmo de relevância binária é uma abordagem de transformação de problema comumente encontrada na literatura que visa separar o problema de classificação multirrótulo em vários problemas binários distintos, que podem então ser usados por algoritmos desenvolvidos para abordar problemas de classificação binária. A Figura 2.3 apresenta uma representação gráfica da classificação por relevância binária, na qual um conjunto de atributos de entrada X é classificado resultando em diversos conjuntos de predição Y, sendo que cada conjunto Y apresenta a predição de um único rótulo.

Figura 2.3: Representação da classificação por Relevância Binária



Fonte: Adaptado de Briggs et al. (2015)

Read et al. (2011) citam que, apesar do método de relevância binária ser muitas vezes desconsiderado na literatura por não extrair correlações diretas entre rótulos presentes na base de dados, esse tipo de abordagem pode resultar em alta performance preditiva sem apresentar problemas de escalabilidade presentes em outros métodos.

Os autores ainda notam as vantagens da separação do problema da classificação multirrótulo em problemas de classificação binária independentes para cada classe incluem a flexibilidade de suportar cenários de dados dinâmicos pois, como a classificação de cada rótulo é completamente independente dos outros, rótulos podem ser incluídos ou removidos sem que o restante do modelo seja afetado. A baixa complexidade computacional da abordagem também é citada, sendo que ela cresce linearmente de acordo com o número de rótulos presentes na base estudada, enquanto que métodos mais complexos podem ter crescimento quadrático. Por fim, Read et al. (2011) citam que a não modelagem de correlações entre rótulos pode afetar negativamente a performance da abordagem, e que outros métodos podem ser criados para abordar tal problema.

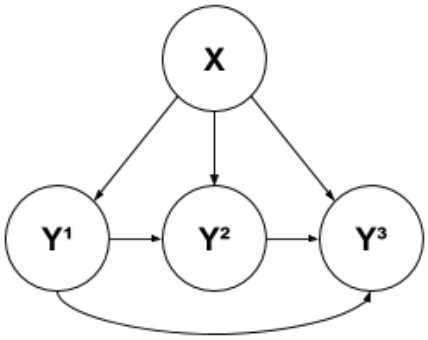
2.4.2 Correntes de Classificadores

A abordagem baseada em Correntes de Classificadores (*Classifier Chains*) proposta por Read et al. (2011) é similar à relevância binária por transformar um problema de classificação multirrótulo em problemas de classificação binária para cada rótulo da base de dados.

A principal diferença proposta pela abordagem de correntes de classificadores é que cada classificador binário tem como entrada não só os atributos da instância a qual está analisando, mas adicionalmente os resultados de classificação preditos por todos os classificadores já executados. Dessa forma, informações sobre correlações entre rótulos são consideradas na criação de cada modelo, já que cada resultado de classificação tem informações sobre todas as classificações anteriores, e terá o resultado de sua classificação propagado para todas as classificações binárias dos rótulos à seguir.

A Figura 2.4 apresenta uma representação gráfica da classificação por Correntes de Classificadores. Pode-se perceber que esta figura é similar àquela apresentada na Figura 2.3, que mostra uma representação do algoritmo de Relevância Binária. A principal diferença para o caso de Correntes de Classificadores é que os conjuntos preditos Y^2 e Y^3 não possuem como entrada somente o conjunto X, mas também recebem como entrada os valores preditos para todos os rótulos anteriores à eles na corrente.

Figura 2.4: Representação da classificação por Correntes de Classificadores



Fonte: Adaptado de Briggs et al. (2015)

Read et al. (2011) notam que a abordagem descrita, utilizando valores binários como atributos utilizados por classificações futuras, não requer procedimentos internos de validação e que a complexidade computacional é idêntica àquela encontrada na abordagem de relevância binária quando o número de rótulos na base é inferior ao número de atributos de cada instância. Por fim, os autores citam que o processo de treinamento das correntes de classificadores pode ser paralelizado, de forma que apenas um único problema de classificação binária esteja ativo

em memória em um dado instante.

2.4.3 RAKEL (*Random k-labelsets*)

O método RAkEL (*Random k-labelsets*) proposto por Tsoumakas e Vlahavas (2007) consiste na divisão aleatória de *labelsets* possíveis presentes na base estudada, abordando então cada subconjunto da base inicial como um problema de rótulo único.

Os autores explicam que, seja $L = \lambda_i$, $i = 1..|L|$ o conjunto de rótulos presentes em um domínio de classificação multirrótulo. Um conjunto $Y \subseteq L$ com $k = |Y|$ é chamado de *k-labelset*. Usando L^k para denotar o conjunto de todos os *k-labelsets* distintos em L. O tamanho de L^k é dado pelo coeficiente binomial $|L^k| = \binom{|L|}{k}$.

O algoritmo RAkEL constrói iterativamente um conjunto de m classificadores *Label Powerset* (LP). Sendo que a cada iteração $i = 1..m$ o *labelset* Y_i é selecionado de L^k sem reposição. O algoritmo então realiza o aprendizado do classificador LP $h_i : X \rightarrow P(Y_i)$. O pseudocódigo da fase de criação de conjuntos de classificadores do algoritmo pode ser observada no Algoritmo 3.

Algoritmo 3: Criação de grupos de classificadores do algoritmo RAkEL

Entrada: Número de modelos m, tamanho do *labelset* k, conjunto de rótulos L, conjunto de treinamento D

Saída: Conjunto de classificadores LP h_i, e seus *k-labelsets* correspondentes Y_i

$R \leftarrow L^k$

para $i = 1 \rightarrow min(m, |L^k|)$ **faça**

- $Y_i \leftarrow$ um *k-labelset* aleatoriamente selecionado de R
- treine um classificador LP $h_i : X \rightarrow P(Y_i)$ em D
- $R \leftarrow R\ Y_i$

fim

Fonte: Adaptado de Tsoumakas e Vlahavas (2007)

Tsoumakas e Vlahavas (2007) notam que valores significativos para o parâmetro de entrada k variam de 2 a $|L| - 1$, pois execuções do algoritmo com parâmetros $k = 1$ e $m = |L|$ resultam na abordagem de classificação de relevância binária, enquanto que parâmetros $k = |L|$ e $m = 1$ resultam na abordagem LP, na qual cada *labelset* da base é considerado como uma classe única para um problema de classificação multiclasse.

Uma vez que todos os grupos de classificadores foram definidos pelo Algoritmo 3, uma nova instância x pode ser classificada. Nesse caso, cada modelo h_i provê decisões

binárias $h_i(x, \lambda_j)$ para cada rótulo λ_j do *k-labelset* Y_i. Após a realização das classificações, o algoritmo calcula a decisão mediana da instância para cada rótulo λ_j em L e dá como saída uma decisão final positiva se tal decisão mediana ultrapassar um limiar t provido pelo usuário. O pseudocódigo contendo a combinação de previsões do algoritmo RAkEL pode ser observado no Algoritmo 4.

Algoritmo 4: Combinação de previsões do algoritmo RAkEL

Entrada: Nova instância x, conjunto de classificadores LP h_i, conjunto de *k-labelsets* Y_i, conjunto de rótulos L

Saída: Vetor de classificação multirrótulo *Resultado*

```
para j = 1 → |L| faça
    Soma_j ← 0
    Votos_j ← 0
fim
para i = 1 → m faça
    para cada rótulo λ_j ∈ Y_i faça
        Soma_j ← Soma_j + hì(x, λ_j)
        Votos_j ← Votos_j + 1
    fim
fim
para j = 1 → |L| faça
    Mediana_j ← Soma_j / Votos_j
    se Mediana_j > t então
        Resultado_j ← 1
    fim
    senão
        Resultado_j ← 0
    fim
fim
```

Fonte: Adaptado de Tsoumakas e Vlahavas (2007)

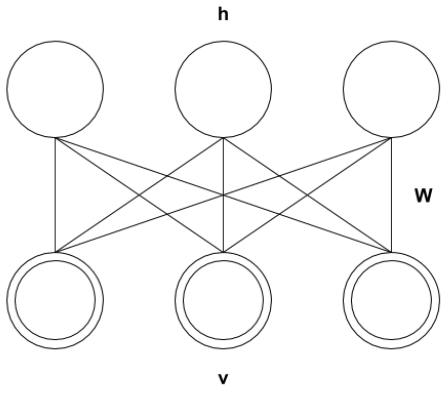
2.4.4 DBPNN (*Deep Back-Propagation Neural Network*)

O algoritmo de classificação DBPNN (*Deep Back-Propagation Neural Network*) basea-se no trabalho publicado por Hinton e Salakhutdinov (2006), no qual a dimensionalidade dos vetores representando características de cada instância de uma base de dados é reduzida usando-se uma rede neural multicamada que tem seus pesos inicializados usando-se de diversas redes RBMs (*Restricted Boltzmann machines*).

Segundo Salakhutdinov e Hinton (2009), BMs (*Boltzmann Machines*) são redes de unidades estocásticas binárias acopladas. A rede é composta de um conjunto de unidades visíveis $v \in \{0,1\}^D$ e um conjunto de unidades ocultas $h \in \{0,1\}^P$, assim como um conjunto de parâmetros $\theta = W, L, J$, em que W, L e J representam respectivamente ligações simétricas ocorrendo entre elementos das camadas visível e oculta, ligações entre elementos das camadas oculta e oculta, e ligações entre elementos das camadas visível e visível, sendo que dada unidade não possui uma ligação com si mesmo.

Já a versão restrita da rede BM, denominada RBM, é uma variação da rede em que ligações entre unidades de uma mesma camada não são permitidas, fazendo com que seus parâmetros L e J não sejam presentes. Segundo os autores, tal restrição faz com que a rede torne-se mais eficiente. Os autores também notam que RBMs podem ser utilizadas para prover data para RBMs em camadas superiores, o que pode levar a um modelo generativo hibrido chamado de *deep belief net* (rede de crenças profundas). A Figura 2.5 apresenta uma representação de uma RBM, notando-se que todas as unidades das camadas visível e oculta são conectadas de forma não-direcional.

Figura 2.5: Representação de uma RBM



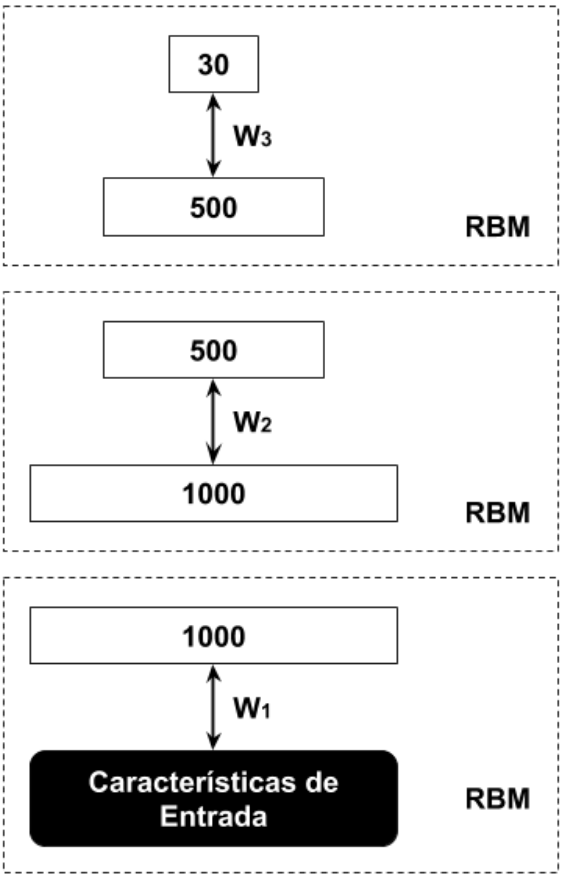
Fonte: Adaptado de Salakhutdinov e Hinton (2009)

Hinton e Salakhutdinov (2006) notam que devido à dificuldade de ajuste de pesos em redes neurais com diversas camadas ocultas, algoritmos que estimem pesos iniciais para a rede podem ser utilizados para que se obtenham melhores resultados.

A técnica descrita por Hinton e Salakhutdinov (2006) envolve a utilização de diversas redes RBMs, que quando usadas sequencialmente podem reduzir a dimensionalidade de uma entrada de dados por empregar uma quantidade decrescente de unidades em sua camadas ocultas. A Figura 2.6 apresenta uma possível configuração, mostrando a quantidade de unidades em cada camada para cada RBM ilustrada, sendo que W_1, W_2 e W_3 representam os pesos

atribuídos a cada ligação entre duas camadas distintas, que por sua vez ligam-se a camadas com quantidade decrescente de unidades.

Figura 2.6: Exemplo de redução da dimensionalidade na rede neural por RBMs

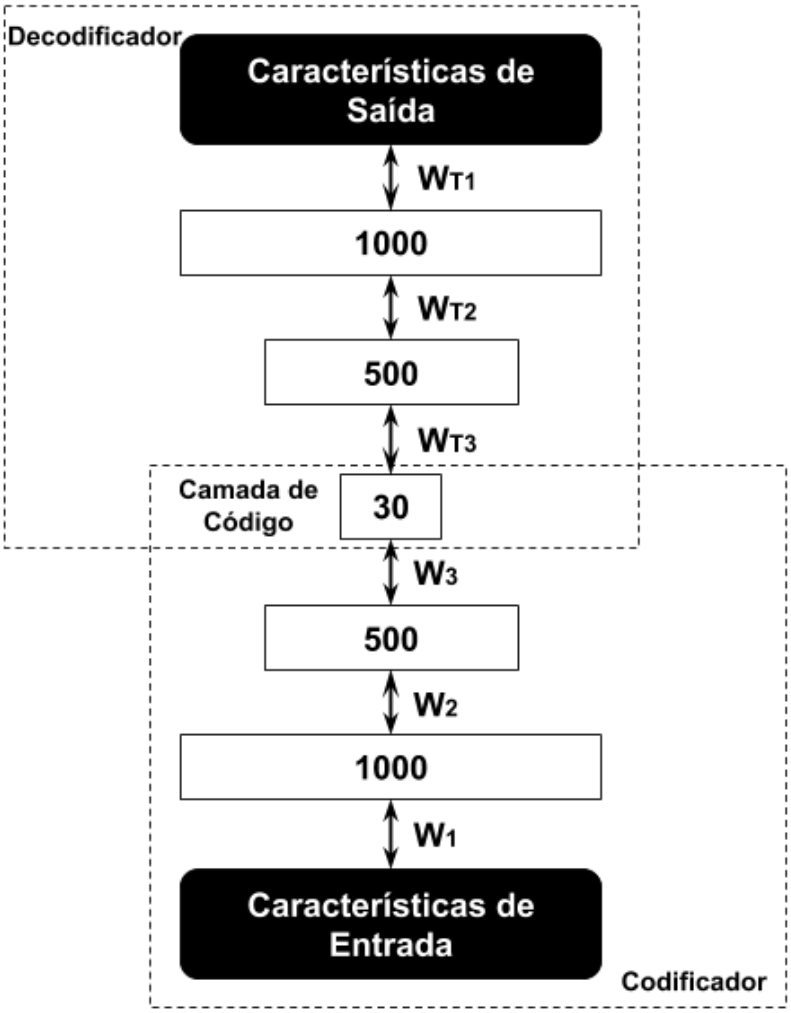


Fonte: Adaptado de Hinton e Salakhutdinov (2006)

Uma vez que cada RBM seja ajustada independentemente, suas camadas podem ser conectadas de forma a construir uma nova rede que represente as entradas de maneira codificada com dimensionalidade reduzida. Pode-se então replicar essa rede de forma reversa como exemplificado pela Figura 2.7, de forma que a rede neural resultante re-codifique as características de entrada em características de saída.

Hinton e Salakhutdinov (2006) demonstram que a inicialização de pesos de uma rede neural por esse método pode facilitar tanto a redução de dimensionalidade para a representação de dados como imagens, como para a construção de redes que possam realizar classificação de instâncias, usando-se da representação intermediária obtida para realizar a classificação de documentos textuais usando-se de contagens de termos utilizados na base de dados como entrada da rede.

Figura 2.7: Exemplo de rede neural construída por encadeamento de RBMs



Fonte: Adaptado de Hinton e Salakhutdinov (2006)

2.5 MODELOS UTILIZADOS NA EXTRAÇÃO DE CARACTERÍSTICAS

Nesta subseção são apresentados os diversos modelos empregados no trabalho atual para representar e extrair características de instâncias textuais de bases de dados.

2.5.1 TF-IDF (*Term Frequency - Inverse Document Frequency*)

Segundo Rajaraman e Ullman (2011), no domínio da categorização automática de documentos é comum que tópicos sejam identificados usando-se do reconhecimento de termos que caracterizem documentos de tal tópico. Os autores notam que um simples ranqueamento de

termos por sua frequência na base de dados não é uma medida eficaz para a identificação de termos característicos, pois tais termos serviriam apenas para construir as sentenças de um documento sem que contenham particular significância em relação ao tópico discutido nele.

Rajaraman e Ullman (2011) também citam que termos raros são mais indicativos de tópicos de documentos mas que esses nem sempre possuem mesma utilidade, sendo que quanto mais frequente o termo é em um documento, mais indicativo do tópico do documento ele seria.

Com base nessas informações, a medida TF-IDF (*Term Frequency - Inverse Document Frequency*) foi elaborada para para representar formalmente a proporção em que um dado termo de uma base de dados está concentrada em um número relativamente pequeno de documentos. A medida assume que em uma coleção de N documentos, $f_i j$ seja o número de ocorrências do termo i no documento j, definindo $TF_i j$ como mostrado na Equação 12.

$$TF_i j = \frac{f_i j}{\max_k f_k j} \tag{12}$$

Em seguida, a frequência inversa de documentos é calculada supondo que um termo i ocorra n_i dos N documentos da base estudada, usando a fórmula presente na Equação 13. Por fim, o valor TF-IDF de um termo i em um documento j é dado pela combinação dos valores $TF_i j$ e IDF_i, segundo a fórmula presente na Equação 14.

$$IDF_i = \log_2(N/n_i) \tag{13}$$

$$TFIDF_i j = TF_i j \times IDF_i \tag{14}$$

O cálculo de valores TF-IDF pode ser ajustado para não só calcular a concentração relativa de termos individuais mas de qualquer n-gramas, combinações entre termos n adjacentes no documento, possíveis. Pode-se também alterar a definição de um documento na base de dados para que, por exemplo, considere-se todas as sinopses pertencentes a um mesmo gênero cinematográfico como fazendo parte de um documento único. Fazendo assim com que $TFIDF_i j$ represente a concentração relativa de n-gramas raros i em sinopses do gênero j.

Entretanto, como valores TF-IDF são calculados para cada termo presente em uma base de dados é interessante que se faça uma seleção entre esses termos, de modo que apenas os termos mais significativos para a distinção entre rótulos sejam usados, afim de reduzir a dimensionalidade dos vetores utilizados pela classificação.

Para tal, pode-se usar de um teste estatístico como o chi-quadrado apresentado por Pearson (1900). Primeiramente constrói-se uma tabela relacionando cada rótulo i presente na base de dados com todos os j possíveis n-gramas extraídos, no qual cada elemento T_{ij} é representado pela soma de todos os valores TF-IDF dos n-gramas j em documentos da classe i. Em seguida, considera-se a distribuição de rótulos presentes no conjunto de treinamento em conjunto com uma distribuição normal para que se encontre k n-gramas com maior probabilidade de serem termos característicos para a base estudada. Ao final da extração, o conjunto de características extraído para cada documento é um vetor de k dimensões apresentando o valor TF-IDF de cada um dos n-gramas selecionados para o documento representado.

2.5.2 LDA (*Latent Dirichlet Allocation*)

O modelo LDA (*Latent Dirichlet Allocation*), apresentado por Blei et al. (2003), é descrito como um modelo generativo probabilístico capaz de modelar relações entre dados discretos como texto.

Os autores definem o modelo capaz de representar documentos (instâncias de uma base de dados) como combinações aleatórias de tópicos latentes, sendo que cada tópico é caracterizado por uma distribuição de termos pertencentes a ele. Uma vez que tais características do modelo são estabelecidas, seu processo generativo para cada documento w de uma base D é dado pelas seguintes etapas:

1. Escolha de $N \sim Poisson(\xi)$.
2. Escolha de $\theta \sim Dir(\alpha)$.
3. Para cada uma das N palavras w_n:
   (a) Escolha de um tópico $z_n \sim Multinomial(\theta)$.
   (b) Escolha de uma palavra w_n de $p(w_n|z_n,\beta)$, a qual é uma probabilidade multinomial condicionada no tópico z_n.

Cada palavra pertence a um vocabulário discreto $1,...,V$, sendo que uma palavra w é dada por um vetor binário V no qual $w^v = 1$ e $w^u = 0$ para todo $u \neq v$. Cada documento é dado por uma sequência de N palavras denotado por $\mathbf{w} = (w_1, w_2, ..., w_N)$. Um córpus é uma coleção de M documentos denotada por $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.

Segundo Faleiros (2016), uma vez que as informações *a priori* α e β, definidas como priores da distribuição de Dirichlet respectivamente relacionadas às distribuições documento-termo e tópico-palavra, podemos descrever a probabilidade das variáveis latentes do modelo: z contendo a representação das distribuições de tópicos de cada palavra para cada documento do córpus; w representando palavras do vocabulário; ϕ vetor contendo a proporção de cada palavra para o vocabulário de cada tópico; e θ representando a distribuição de todos os tópicos entre os documentos. Tal probabilidade é dada pela Equação 15.

$$p(z,w,\phi,\theta|\alpha,\beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \Big(\prod_{i=1}^{V} p(z_{j,i}|\vec{\theta}_j) p(w_{i,j}|z_{i,j},\phi_{z_{i,j}}) \Big) \tag{15}$$

Segundo Faleiros (2016), pode-se usar o teorema de Bayes para formular $p(z,w,\phi,\theta|\alpha,\beta)$ como o cálculo *a posteriori* do LDA, resultando na fórmula apresentada na Equação 16.

$$p(z,\phi,\theta|w,\alpha,\beta) = \frac{p(z,w,\phi,\theta|\alpha,\beta)}{p(w)} \tag{16}$$

Faleiros (2016) nota que a Equação 16 mostra um processo que pode ser considerado como o inverso do processo generativo apresentado pelo modelo LDA que pode ser usado para gerar a distribuição *a priori* do modelo. O autor cita que, embora o processo seja muito custoso computacionalmente, há alguns métodos de aproximação das probabilidades na literatura, como *Gibbs Sampling* (Amostrador de Gibbs) e *Variational Inference* (Inferência Variacional).

2.5.3 *Word Embeddings*

Segundo Le e Mikolov (2014), muitos algoritmos de aprendizado de máquina requerem como entrada vetores de característica de tamanho fixo, mas abordagens comumente utilizadas para a representação de características textuais usam modelos *bag-of-words* (sacolas de palavras) que não levam em consideração a ordem ou significado semântico das palavras que representam.

Um dos conceitos apresentado pelos autores para abordar esse problema é o de vetores distribuídos de representação de palavras, no qual a tarefa principal é a de predizer uma termo à partir de uma coleção de palavras em um contexto. Os autores notam que nesse modelo cada palavra da base de dados é representada por uma coluna em uma matriz W, indexada pela posição da palavra no vocabulário da base.

Le e Mikolov (2014) apresentam na definição formal do modelo que dada uma

sequência de palavras de treinamento $w_1, w_2, w_3, ..., w_T$, o objetivo é maximizar a probabilidade logarítmica presente na Equação 17.

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t|w_{t-k},...,w_{t+k}) \tag{17}$$

A tarefa de predição é então realizada por um classificador multiclasse, o qual os autores citam como comumente sendo *softmax*. O cálculo das probabilidades pode ser então descrito como na Equação 18, sendo que cada y_i é dado pela probabilidade logarítmica não normalizada de cada palavra de saída i como demonstrado na Equação 19, na qual U e b são os parâmetros de *softmax*, e h é construído da concatenação ou média dos vetores de palavra de W.

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{18}$$

$$y = b + Uh(w_{t-k},...,w_{t+k};W) \tag{19}$$

Le e Mikolov (2014) também apresentam um outro modelo baseado no anterior, no qual o contexto presente no parágrafo em que cada termo está presente também é considerado na construção do modelo. Nesse modelo, denominado de *Paragraph Vector* (Vetor de Parágrafos), Le e Mikolov (2014) explicam que cada parágrafo contido na base de dados é representado por uma coluna em uma matriz D, sendo que cada palavra é mais uma vez representada por uma coluna em uma matriz W.

Os autores notam que a única alteração formal deste modelo em comparação ao apresentado anteriormente está presente na Equação 19, sendo que h deve ser construído levando tanto W como D em consideração. Os autores explicam que um indicador do parágrafo analisado é incluído no contexto de cada termo do conjunto de treinamento, considerado como um termo adicional capaz de manter informações sobre o contexto ao qual cada palavra pertence. As duas etapas principais da abordagem podem então ser descritas como:

1. Treinamento, adquirindo vetores W de palavras, parâmetros *softmax* U e b e vetores de parágrafo D.

2. Inferência de novos vetores de parágrafo, adicionando mais colunas em D, usando a descida no gradiente em D enquanto valores em W, U e b permanecem fixos, e uso de D para fazer predições acerca de rótulos usando-se de classificadores tradicionais.

Por fim, Le e Mikolov (2014) citam que *paragraph vectors* podem ser utilizados diretamente como características de entrada de algoritmos de aprendizado de máquina convencionais, notando que uma vantagem apresentada pela abordagem é a de ter seu aprendizado à partir de dados não rotulados, tendo boa performance para tarefas nas quais não se tem rótulos suficientes atribuídos à base estudada.

Nota-se que outros modelos utilizados para a representação de palavras como vetores de características podem ainda ser encontrados na literatura, como aqueles apresentados por Kim et al. (2016) no qual o modelo é capaz de fazer predições baseadas apenas nos caracteres que compõem um determinado documento, e Peters et al. (2018) o qual inclui características do contexto linguístico no qual cada palavra está inserida em conjunto com características sintáticas e semânticas em sua modelagem.

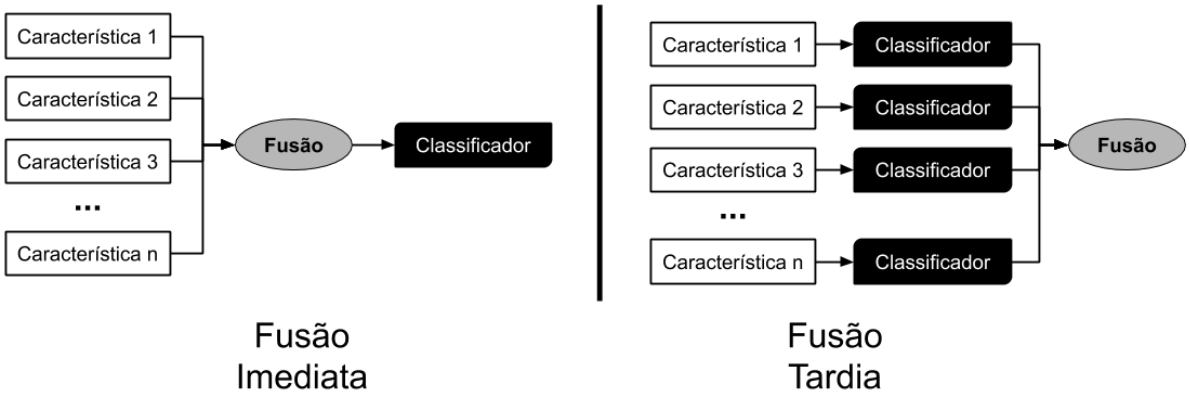
2.6 FUSÃO IMEDIATA E FUSÃO TARDIA

Para que se possa realizar a combinação de diversas características de forma que possam ser utilizadas na mesma classificação, abordagens de fusão imediata (*early fusion*) e fusão tardia (*late fusion*) podem ser exploradas.

A Figura 2.8 apresenta representações gráficas do funcionamento de abordagens de fusão imediata e fusão tardia. Segundo Ebersbach et al. (2017) a fusão imediata é uma fusão à nível de características, na qual vetores de características provenientes de diversas fontes são concatenados em um único vetor que é então utilizado como entrada do classificador estudado. Os autores explicam que o uso de tal vetor, que muitas vezes possui uma grande dimensionalidade, proporciona um aumento nos tempos de treino e classificação do experimento, mas que também pode proporcionar um aumento na performance da classificação.

Ainda segundo Ebersbach et al. (2017) a abordagem de fusão tardia implica em uma fusão à nível de decisão, no qual o processo consiste em realizar a classificação de cada vetor de características individualmente no mesmo classificador e, em seguida, utilizar os resultados obtidos de cada vetor de características em conjunto com determinadas estratégias de combinação para que se obtenha uma única classificação final. Estratégias de combinação encontradas na literatura podem fazer uso da medida de confiança que o classificador atribui para cada classificação, fazendo a combinação entre os valores obtidos para cada vetor de característica.

Figura 2.8: Representações de abordagens de fusão imediata e fusão tardia



Fonte: Adaptado de Ebersbach et al. (2017)

3

TRABALHOS RELACIONADOS

Nesta seção são apresentados os trabalhos presentes na literatura que possuem contribuições relevantes ao presente estudo. Trabalhos relacionados ao problema da classificação de gêneros cinematográficos são explorados afim de prover uma visão mais ampla do estado da arte do domínio estudado. Um conjunto de trabalhos diretamente relacionados, devido ao fato de apresentarem estudos relacionados à classificação multirrótulo de gêneros cinematográficos usando características extraídas de suas sinopses, também é apresentado assim como suas abordagens e resultados, que podem ser usados como comparações mais diretas ao presente trabalho.

Trabalhos como os de Rasheed et al. (2005), Zhou et al. (2010), Huang e Wang (2012), Sugano et al. (2003) e Wehrmann e Barros (2017) usam a estratégia frequentemente encontrada na literatura de abordar a classificação de filmes usando-se de características audiovisuais muitas vezes extraídas de seus *trailers*. Outros tipos de características também encontrados na literatura são utilizados por estudos como os de Austin et al. (2010), que utilizou características extraídas da trilha sonora de filmes como um todo para fazer a classificação, e de Ivasic-Kos et al. (2014), que realizou a classificação de gêneros cinematográficos usando-se de características visuais extraídas de pôsteres de filmes.

Rasheed et al. (2005) apresentam uma estratégia de classificação de filmes baseada em características visuais de seus *previews*. Informações como: Movimentação de câmera; Variação de cor da cena; Duração e Contraste das cenas são analisadas para a construção de um classificador. Os autores descrevem o processo de treinamento do classificador utilizando a técnica de *Mean Shift* para a clusterização do conjunto de treinamento no espaço 4-dimensional estudado. Os autores observam que não é intuitivo que gêneros cinematográficos sejam

classificados binariamente, mas que idealmente filmes são classificados como combinações de gêneros. Resultados obtidos mostraram que em um corpus de 101 *previews* de filmes, 17 apresentaram-se como *outliers* da classificação. Os autores afirmam que, enquanto esse número não pode ser visto como acurácia de 87%, mas sim como uma forte evidência de que o mapeamento de filmes em características de baixo nível de vídeo possa corresponder ao mapeamento existente entre os gêneros cinematográficos existentes.

Zhou et al. (2010) apresentam um método para a categorização de gêneros cinematográficos a partir de características extraídas de cenas de seus *trailers*. Cada *trailer* é decomposto em um conjunto de cenas por meio da análise das fronteiras visuais que separam uma cena da outra. Após esse passo, o quadro central de cada cena é extraído para sua classificação. Em seguida, características do quadro analisado são extraídas e seus descritores são usados pelo algoritmo *K-means*, com K = 100, para que os *trailers* sejam agrupados em *clusters*. Os experimentos descritos no trabalho classificaram 1239 *trailers* nas categorias: ação, comédia, drama, e horror. Os autores relatam que a acurácia média entre os experimentos ficou em torno de 70%, com o melhor resultado apresentando uma acurácia de $74,7\%$.

Huang e Wang (2012) propõem o uso da metaheurística *Self-Adaptive Harmony Search* (SAHS) para escolher um subconjunto de características dentro de 277 possíveis características extraídas do áudio e vídeo de *trailers* de filmes para fazer sua classificação em gêneros. Segundo os autores, uma vez que um subconjunto de características seja escolhido, um classificador SVM (*Support Vector Machine*) é utilizado e, por meio da escolha majoritária, o gênero é atribuído ao *trailer* analisado. Os experimentos foram realizados com um conjunto de 223 *trailers* de filmes do *website Apple Movie Trailers*. A distribuição de gêneros cinematográficos na base estudada foi: 35 de ação, 14 de animação, 46 de comédia, 28 para documentário, 67 de drama, 7 de musical e 26 para suspense. Os resultados mostraram uma acurácia de até $91,9\%$.

Sugano et al. (2003) apresentam um método de classificação de gêneros cinematográficos baseado na análise de características audiovisuais de cenas extraídas de filmes. Os autores extraíram um total de 7450 cenas provenientes de 4 filmes distintos. Em seguida um conjunto de 347 cenas aleatórias foi escolhido para o treinamento de um classificador LMDT (*Linear Machine Decision Tree classifier*). O treinamento do classificador foi realizado usando-se de 8 vetores de características extraídos de cada cena, características extraídas incluíram: duração da cena, natureza do áudio da cena (silêncio, diálogo, outros) e quantidade de movimento detectado na cena. As cenas foram classificadas manualmente como: cena de ação, cena dramática, conversação e genérica. Os resultados compilados por filme mostram acurácia de até 93%.

Wehrmann e Barros (2017) apresentam resultados de um classificador baseado em redes neurais convolucionais (*ConvNets*) na classificação multirrótulo de gêneros cinematográficos usando-se de características visuais extraídas de *trailers* de filmes. A base de dados usada no estudo foi *Labelled Movie Trailer Dataset*, que contém cerca de 10000 *trailers* de filmes classificados em 22 gêneros diferentes. Os dados foram pré-processados de forma que 9 gêneros foram escolhidos para o estudo: ação, aventura, comédia, crime, drama, horror, romance, ficção científica e suspense, sendo que os conjuntos de treinamento e teste utilizados possuíram um total de 2861 e 772 entradas respectivamente. Segundo os autores, o classificador construído mostra resultados melhores que aqueles presentes no estado da arte da classificação usando um único rótulo de gêneros cinematográficos.

Austin et al. (2010) propõem que trilhas sonoras de filmes podem conter informações importantes para sua classificação. Os autores construíram uma base de dados de 98 trilhas sonoras de filmes, classificadas da seguinte forma: 25 como romance, 25 como drama, 23 como horror, e 25 como ação. Para a realização do estudo, um total de 1728 músicas foram obtidas a partir dos filmes da base de dados. Os 60 segundos no meio de cada música foram então usados para a extração de características de áudio, que por sua vez foram usadas para o treinamento de um classificador SVM (*Support Vector Machine*). O estudo apresenta resultados de acurácia que variam de 60% para a classe ação, até $43,5\%$ para a classe horror. Os autores notam que a baixa taxa de acerto em alguns gêneros pode surgir pela semelhança entre características musicais desses gêneros, tais como os gêneros romance e drama que possuem muitas faixas musicais semelhantes.

Ivasic-Kos et al. (2014) apresentam resultados de classificadores multirrótulo treinados para identificar gêneros cinematográficos a partir de pôsteres usados para promover seus filmes. A base de dados escolhida para o estudo era composta de 6739 imagens classificadas com 18 gêneros: ação, aventura, animação, comédia, crime, desastre, documentário, drama, fantasia, história, horror, mistério, romance, ficção científica, suspense, *thriller*, guerra e *western*, sendo que, após pré-processamento da base, esse número foi condensado em 11 gêneros através da junção de gêneros que os autores julgaram similares. Em seguida características de baixo nível, como histogramas de cor, valor e saturação, são extraídas das imagens e usadas para treinamento dos classificadores. Três classificadores foram construídos no estudo: *Naive Bayes*, no qual o problema da classificação multirrótulo foi reduzido a vários problemas de classificação em um único rótulo; *Random k-label sets* (RAKEL); e *Multi-Label K-Nearest Neighbor* (ML-kNN). Resultados relatados pelos autores mostram que a medida-F do melhor caso dos classificadores foi 38%.

Mesmo que os resultados desses trabalhos não possam ser diretamente comparados com os deste estudo por abordarem problemas de classificação distintos, a composição de suas bases de dados e os métodos abordados em suas classificações provam-se interessantes para a composição deste trabalho. O Quadro 3.1 apresenta um resumo das características mais relevantes dos trabalhos apresentados acima, em respeito ao algoritmos de classificação implementados, natureza das características de suas bases de dados, número de gêneros cinematográficos considerados para classificação, e tipo da classificação realizada.

Quadro 3.1: Resumo dos estudos relacionados à este trabalho

Estudo	**Método de Classificação**	**Natureza das Características**	**Gêneros Considerados**	**Tipo da Classificação**
Rasheed et al. (2005)	*Mean Shift*	Visuais	4	Rótulo Único
Zhou et al. (2010)	*K-means*	Visuais	4	Rótulo Único
Huang e Wang (2012)	SVM	Audiovisuais	7	Rótulo Único
Sugano et al. (2003)	LMDT	Audiovisuais	4	Rótulo Único
Wehrmann e Barros (2017)	*ConvNets*	Visuais	9	Multirrótulo
Austin et al. (2010)	SVM	Trilhas Sonoras	4	Rótulo Único
Ivasic-Kos et al. (2014)	ML-kNN, RAKEL, *Naive Bayes*	Visuais	11	Multirrótulo

Fonte: Autoria Própria

Entre os trabalhos encontrados na literatura, um conjunto pode ser considerado diretamente relacionado ao tema abordado neste estudo. Publicações como as de Hoang (2018), Rahman et al. (2017) e Ho (2011) abordam a classificação de gêneros cinematográficos usando-se de características extraídas de suas sinopses. No entanto, como abordaremos em seguida, pode-se notar que entre esses estudos não há um grande foco na combinação de características provenientes de métodos distintos de extração, ou no estudo de características na língua portuguesa especificamente como é o caso do estudo descrito por este documento.

Hoang (2018) aborda o problema da classificação multirrótulo de gêneros cinematográficos usando-se de três métodos distintos de classificação: *Naive Bayes*; *Word2Vec+XGBoost*; e Redes Neurais Recorrentes, aplicados a uma base de dados composta por 255.853 sinopses em inglês extraídas do *website* IMDB (*Internet Movie Database*). As características utilizadas no estudo incluem representações *bag of words* e *word embeddings*. O autor nota que ao fazer uso de *Gated Recurrent Units* (GRU), o melhor caso de classificação alcançou resultados de até $0,56$ de medida-F, 50% de Índice de Jaccard e $80,5\%$ de *hit rate*.

Rahman et al. (2017) apresentam um estudo focado especificamente na classificação multirrótulo de sinopses de filmes indianos. Os autores conduziram experimentos com algoritmos de aprendizagem como *Naive Bayes*, regressão logística, *K-Nearest Neighbors* (KNN), árvores de decisão e *Support Vector Machines* (SVM) lineares em uma base de dados contendo 13.868 sinopses de filmes de origem indiana. Características utilizadas no estudo são reportadas como utilizando vetores de ocorrência das palavras da base estudada. O melhor resultado reportado pelo estudo conta com uma precisão média de $0,421$, revocação média de $0,36$ e medida-F média de $0,386$ usando-se do algoritmo *Naive Bayes*.

Ho (2011) estuda o problema da classificação multirrótulo de sinopses de filmes analisando uma base de dados contendo 16.000 sinopses em inglês extraídas do *website* IMDB. O estudo analisou o desempenho de quatro métodos de classificação distintos: Abordagem *One-Vs-All* usando SVM; *Multi-label K-nearest neighbors* (ML-KNN); *Parametric Mixture Model* (PMM); e Redes Neurais. Características utilizadas no estudo foram compostas de uma representação *bag of words* dos termos de cada instância da base. Os melhores resultados divulgados pelo autor contam com uma precisão de $0,51205$, revocação de $0,61631$ e medida-F de $0,54999$, usando-se do classificador SVM treinado em um subconjunto balanceado do conjunto inicial de treinamento.

O Quadro 3.2 apresenta um resumo sobre as principais características que podem ser observadas nos estudos diretamente relacionados à este trabalho, citando se houve uma seleção da base de dados quanto à sua origem ou linguagem em que foi escrita, número de sinopses presentes na base estudada, métodos de classificação empregados no estudo, e melhores resultados obtidos ao final do estudo.

Quadro 3.2: Resumo dos estudos diretamente relacionados à este trabalho

Estudo	Composição das Sinopses	Tamanho da Base de Dados	Métodos de Classificação	Melhores Resultados
Hoang (2018)	Inglesas	$255,853$	*Naive Bayes*; *Word2Vec+XGBoost*; Redes Neurais Recorrentes	$0,56$ Medida-F; 50% Índice de Jaccard; $80,5\%$ *Hit rate*
Rahman et al. (2017)	Indianas	$13,868$	*Naive Bayes*; Regressão Logística; KNN; Árvores de Decisão; SVM	$0,421$ Precisão; $0,36$ Revocação; $0,386$ Medida-F
Ho (2011)	Inglesas	$16,000$	SVM (*One-Vs-All*); ML-KNN; PMM; Redes Neurais	$0,51205$ Precisão; 0.61631 Revocação; 0.54999 Medida-F

Fonte: Autoria Própria

4

DESENVOLVIMENTO

Nesta seção são descritas todas as informações relevantes à construção dos experimentos que foram realizados neste trabalho. Descrições de todas as bases de dados extraídas para este trabalho são apresentadas, assim como experimentos preliminares realizados com anotadores humanos aplicados a elas. Em seguida é descrita a configuração dos demais experimentos realizados com algoritmos classificadores, apresentando a lista completa de características textuais extraídas de todas as entradas da base de dados e, por fim, todos os classificadores utilizados para a realização de experimentos.

4.1 A BASE P-TMDB

Nesta seção a base de dados P-TMDb, uma base composta por sinopses de filmes na língua portuguesa extraída do *website The Movie Database* é apresentada. Após uma breve introdução sobre o *website* do qual a base foi extraída é feita uma discussão sobre as principais características da base de dados, assim como sobre a construção de duas bases derivadas, P-TMDb(-) e P-TMDb(+), que foram obtidas após o rebalanceamento da base original.

4.1.1 Sobre o TMDb

O *The Movie Database (TMDb)*[1] é um projeto estabelecido em 2008 dedicado ao arquivamento e distribuição de metadados de filmes e séries de televisão. O *website* conta atualmente com informações sobre mais de 450.000 filmes e 80.000 séries, incluindo adicionalmente

[1] `https://www.themoviedb.org`

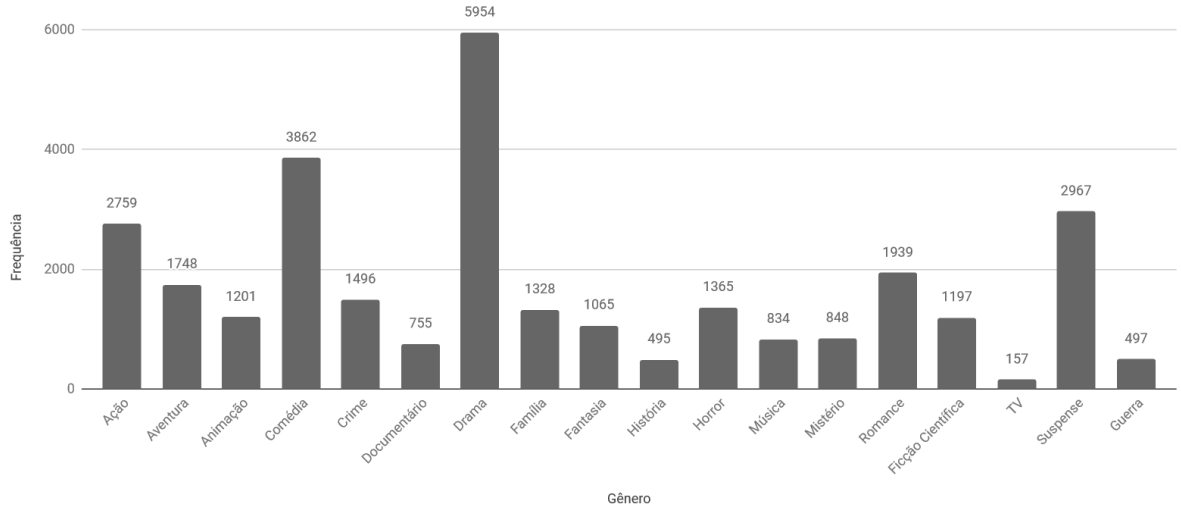
informações sobre suas mais de 100.000 temporadas e mais de 1.600.000 episódios. Informações armazenadas pelo *website* estão disponíveis em 39 línguas diferentes e são acessadas por mais de 200.000 desenvolvedores e companhias por meio de mais de três bilhões de requisições diárias.

4.1.2 Características da Base P-TMDb

Neste estudo foi utilizada a *API* do TMDb para obter títulos, sinopses e rótulos de gêneros de um total de 13.394 filmes, correspondendo ao subconjunto de sinopses em português disponíveis no *website* em Dezembro de 2017. Essa base de dados, nomeada P-TMDb, foi então o foco principal do estudo.

As sinopses da base P-TMDb apresentam 18 gêneros distintos, sendo que uma única sinopse poderia estar rotulada com qualquer combinação desses gêneros. Os gêneros presentes na base são: Ação, Aventura, Animação, Comédia, Crime, Documentário, Drama, Família, Fantasia, História, Horror, Música, Mistério, Romance, Ficção científica, TV, Suspense e Guerra. A frequência de cada gênero individualmente na base de dados coletada pode ser observada na Figura 4.1.

Figura 4.1: Frequência de cada gênero na base P-TMDb



Fonte: Autoria Própria

A Tabela 4.1 apresenta as principais características da base P-TMDb. Observa-se que, mesmo que a base de dados possua 18 rótulos de gêneros distintos, segundo a medida P_{min} cerca de 30% das sinopses está rotulada com um único gênero, sendo que a cardinalidade indica que a base como um todo possui uma média de $2,275$ rótulos ativos por sinopse. Pode-se

também constatar, como observado na distribuição de frequências das Figura 4.1, um claro desbalanceamento entre a frequência de cada gênero individualmente, sendo que em média cada gênero é aproximadamente $6,973$ vezes mais infrequente que o gênero dominante segundo o valor da medida *MeanIR*, e que a proporção entre os gêneros mais e menos frequentes chega a $37,924$ vezes segundo o valor da medida *MaxIR*.

Tabela 4.1: Principais características da base P-TMDb

	P-TMDb
Sinopses	13.394
Gêneros	18
Cardinalidade	$2,275$
Densidade	$0,126$
P_{min}	$0,302$
MaxIR	$37,924$
MeanIR	$6,973$
CVIR	$1,196$

Fonte: Autoria Própria

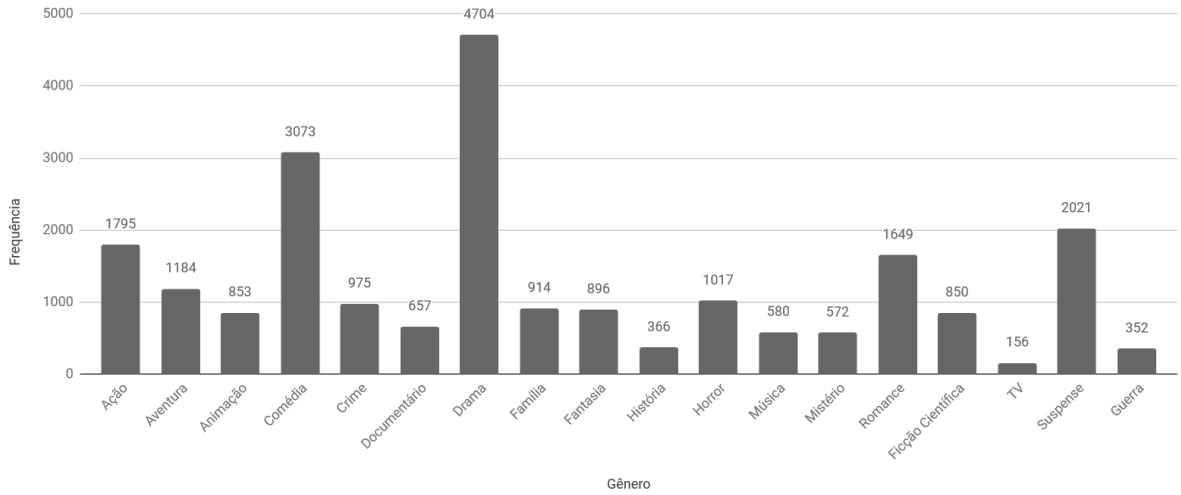
4.1.3 Bases Derivadas por Reamostragem

Devido ao desbalanceamento observado na base P-TMDb e aos impactos que esse fator poderia ter nos resultados de classificação, investigou-se métodos de reamostragem que pudessem ser utilizados para o rebalanceamento da base original.

Após os estudos dos métodos disponíveis, optou-se pelo uso dos algoritmos de reamostragem LP-RUS (*Label Powerset Random Undersampling*) e LP-ROS (*Label Powerset Random Oversampling*), ambos descritos por Charte et al. (2015). Tais algoritmos foram escolhidos por realizarem o rebalanceamento da base original por duplicação de instâncias, não realizando criação de instâncias artificiais. O uso dos algoritmos proporcionou a criação de duas novas bases, chamadas de P-TMDb(-) e P-TMDb(+), nas quais, respectivamente, os *labelsets* mais comuns foram removidos ou os *labelsets* mais raros foram duplicados, para que se obtivesse um decréscimo ou acréscimo de 25% em relação ao número de sinopses da base P-TMDb original. As distribuições de gêneros nas duas bases derivadas podem ser visualizadas nas Figuras 4.2 e 4.3.

Na Figura 4.2 pode-se observar mudanças quanto às proporções de gêneros quando comparadas às observadas na Figura 4.1. As maiores alterações podem ser observadas para os gêneros Suspense e Ação que, apesar de ainda ocuparem a terceira e quarta posição como

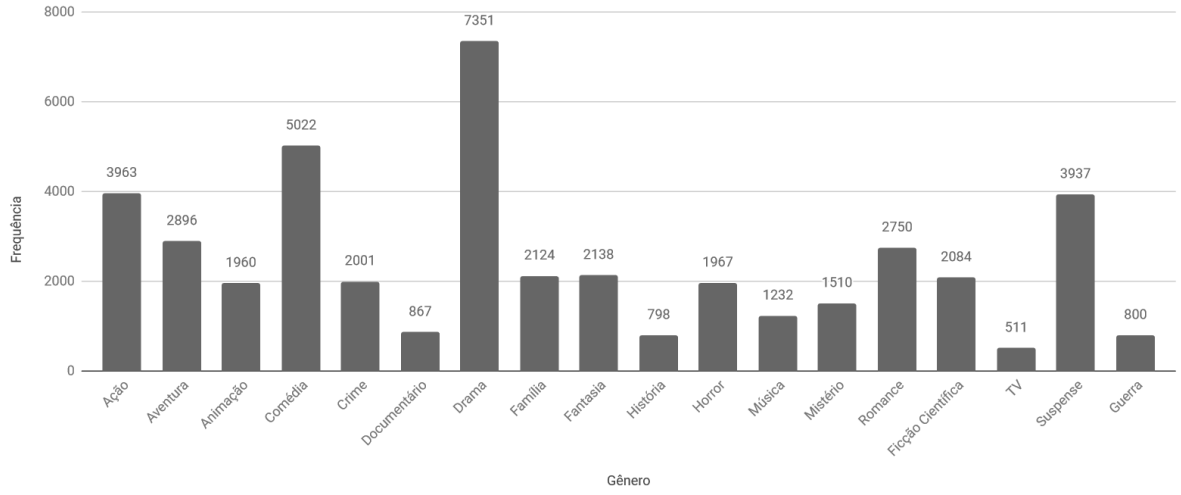
Figura 4.2: Frequência de cada gênero na base P-TMDb(-)



Fonte: Autoria Própria

gêneros mais frequentes, encontram-se com distribuições um pouco mais similares a Romance na base rebalanceada. Pode-se também observar que a execução do algoritmo de reamostragem fez com que as frequências dos gêneros Família e Fantasia se tornassem similares, sendo que anteriormente havia uma maior diferença entre os dois. Por fim, nota-se que, apesar do rebalanceamento remover uma grande quantidade de sinopses pertencentes aos dois gêneros mais frequentes, Drama e Comédia ainda se apresentam como um viés na base P-TMDb(-).

Figura 4.3: Frequência de cada gênero na base P-TMDb(+)



Fonte: Autoria Própria

Já na Figura 4.3 é possível notar mudanças mais significativas quanto às proporções relativas entre os gêneros, sendo imediatamente aparente que todos os gêneros, exceto Drama,

Comédia, Ação e Suspense, apresentam proporções significativamente mais próximas em relação aos quatro gêneros mais frequentes na base. Esse fato é demonstrado numericamente na Tabela 4.2, que mostra a comparação entre as métricas das bases derivadas e a base P-TMDb original.

Tabela 4.2: Comparação entre as características das bases P-TMDb, P-TMDb(-) e P-TMDb(+)

	P-TMDb	P-TMDb(-)	P-TMDb(+)
Sinopses	13.394	10.150	16.803
Gêneros	18	18	18
Cardinalidade	2,275	2,228	2,613
Densidade	0,126	0,124	0,145
P_{min}	0,302	0,355	0,244
MaxIR	37,924	30,154	14,386
MeanIR	6,973	6,948	4,727
CVIR	1,196	0,967	0,738

Fonte: Autoria Própria

A Tabela 4.2 mostra que, após o rebalanceamento da base P-TMDb, o número médio de rótulos ativos ao mesmo tempo não sofreu grandes alterações. As cardinalidades das três bases diferem de maneira tênue, enquanto que a proporção de sinopses rotuladas com apenas um único gênero foi alterada em cerca de cinco pontos percentuais em ambas as variações da base. Esses fatos indicam que as bases derivadas por reamostragem possuem perfis similares quanto à alocação de gêneros por sinopse, e que os algoritmos de rebalanceamento alteraram outras propriedades da base original.

Mudanças mais significativas podem ser observadas quando analisamos as métricas de desbalanceamento das bases como *MaxIR*, *MeanIR* e *CVIR*. Em comparação com a base original, a base P-TMDb(-) possui uma mudança significativa em relação à proporção de desbalanceamento máxima entre seus gêneros, mostrando que o algoritmo LP-RUS conseguiu afetar a distribuição de gêneros na base. No entanto, ao analisar a média entre todas as proporções de desbalanceamento, podemos constatar que a alteração de proporções apresentada entre os gêneros mais e menos frequentes não foi sentida no restante da base. Pode-se então concluir que o algoritmo teve maior impacto no desbalanceamento dos gêneros menos frequentes, não tendo outras alterações significativas em relação à base original.

Entretanto, ao se analisar as métricas que dizem respeito à base P-TMDb(+) pode-se constatar uma grande mudança quanto as proporções de desbalanceamento. Uma redução de mais de 62% da proporção de desbalanceamento máxima e 32% da proporção média mostram que o algoritmo LP-ROS teve sucesso em tratar do desbalanceamento entre os gêneros da base

P-TMDb.

Todas as bases descritas nesta seção foram utilizadas na realização de experimentos com os algoritmos classificadores para que pudessem ser feitas comparações entre os grupos de características e algoritmos estudados neste trabalho.

4.2 EXPERIMENTOS COM ANOTADORES HUMANOS

Para que se possa melhor avaliar os resultados obtidos usando os classificadores e conjuntos de características a serem apresentados neste trabalho é interessante que classificações alternativas sejam realizadas na mesma base de dados.

Na Seção 3 deste trabalho foram apresentados os resultados de trabalhos relacionados que, no geral, utilizaram bases de dados diferentes da utilizada neste trabalho e classificadores induzidos por algoritmos de aprendizado de máquina.

Para que se tivesse uma estimativa do desempenho esperado pela classificação humana no contexto da base de dados estudada, foi conduzido um experimento com anotadores humanos. Os resultados obtidos nesse experimento serviram então como uma base de comparação adicional para a análise dos resultados obtidos ao final deste trabalho.

No experimento descrito à seguir, três anotadores não especialistas na área realizaram a classificação de um subconjunto aleatório de sinopses da base P-TMDb. Os anotadores não se comunicaram durante o experimento e foram instruídos à classificar cada sinopse com todos os gêneros que julgassem apropriados. Para a rotulação, os anotadores receberam planilhas contendo a lista de sinopses que deveriam ser classificadas em conjunto com todos os possíveis rótulos presentes na base, todos os anotadores foram instruídos a levar em consideração somente informações contidas na sinopse, desconsiderando possíveis associações com nomes próprios de atores ou personagens que poderiam afetar a classificação.

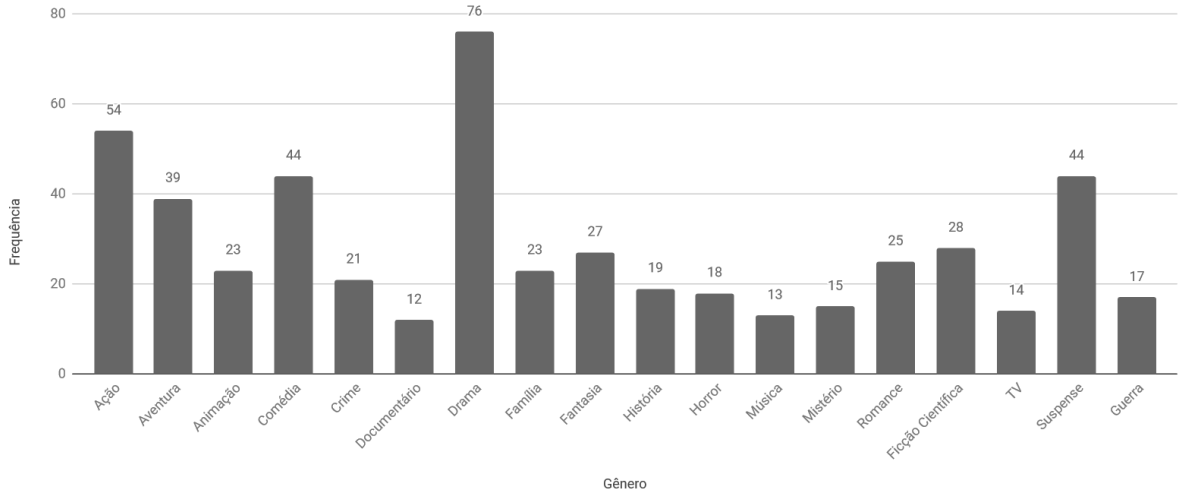
Ao fim do experimento, os resultados de todos os anotadores foram compilados, assim como métricas adicionais aferidas para analisar em quais gêneros específicos o maior índice e tipo de concordância entre os classificadores foi obtido.

A seguir são descritas as características da base P-TMDb(H) utilizada no experimento de anotação, os resultados obtidos ao final do mesmo e uma discussão sobre peculiaridades observadas na anotação humana que podem afetar o desempenho de classificadores automáticos no domínio estudado.

4.2.1 A Base P-TMDb(H)

A base P-TMDb(H) é um subconjunto da base P-TMDb composto por 180 sinopses extraídas aleatoriamente da base original, sendo que 10 sinopses foram extraídas para cada gênero. A Figura 4.4 mostra a frequência de ocorrência dos gêneros na base P-TMDb(H). Nota-se que, mesmo que a extração desse subconjunto tenha sido feita de forma a garantir uma distribuição mais uniforme quando comparada a base original, pode-se perceber um claro desbalanceamento à favor das classes que também mostram-se como sendo mais comuns na base P-TMDb.

Figura 4.4: Frequência de cada gênero na base P-TMDb(H)



Fonte: Autoria Própria

A Tabela 4.3 mostra a comparação entre as características da base P-TMDb(H) e da base P-TMDb original. Pode-se notar que os valores associados à cardinalidade e densidade estão próximos dos valores originais, indicando que de um modo geral a quantidade de rótulos atribuídos a cada sinopse permaneceu similar em ambas as bases. Diferenças mais significativas entre as bases se apresentam nas medidas que representam o número de entradas com apenas um rótulo atribuído e as medidas que descrevem o balanceamento dos rótulos.

Como todas as medidas que apresentam diferenças significativas representam valores atrelados à distribuição dos rótulos da base, os quais poderiam afetar o treinamento de novos modelos de classificação mas em teoria não afetariam os classificadores humanos que realizam tal classificação com modelos já construídos por experiência prévia, a base P-TMDb(H) foi julgada apropriada para os testes com anotadores humanos. Assumiu-se que esses anotadores são classificadores pré-treinados com informações externas, representando o conhecimento sobre a classificação de gêneros cinematográficos que um ser humano "médio" teria.

Tabela 4.3: Comparação entre as características das bases P-TMDb e P-TMDb(H)

	P-TMDb	P-TMDb(H)
Sinopses	13.394	180
Gêneros	18	18
Cardinalidade	2,275	2,844
Densidade	0,126	0,158
P_{min}	0,302	0,117
MaxIR	37,924	6,333
MeanIR	6,973	3,443
CVIR	1,196	0,457

Fonte: Autoria Própria

4.2.2 Resultados da Anotação

Após a realização da classificação pelos anotadores humanos, métricas de análise foram obtidas comparando-se com os rótulos atribuídos no experimento com os originais da base. A Tabela 4.4 apresenta os principais resultados obtidos pelos anotadores individuais, assim como a média simples de todos os resultados.

Tabela 4.4: Resultados da classificação por anotadores humanos

	Anotador 1	Anotador 2	Anotador 3	Média
Acurácia	39,96%	45,25%	52,37%	45,86%
Precisão	50,39%	54,49%	57,81%	54,23%
Revocação	55,84%	63,41%	73,09%	64,11%
Medida-F	52,98%	58,61%	64,56%	58,72%

Fonte: Autoria Própria

Pode-se perceber que, apesar de variações entre os resultados individuais dos classificadores existirem, é possível que se observem tendências nos valores dos resultados obtidos. Os valores de medida-F, que geralmente são utilizados para definir a qualidade da classificação de um modo geral, têm suas médias próximas a 58% enquanto que a acurácia, usada para representar o número de rótulos corretamente atribuídos em proporção do número de rótulos ativos para cada instância, apresenta uma média de 45,86%.

Com o interesse de extrair resultados mais aprofundados sobre a classificação humana, foi feita uma análise dos "equívocos" e acertos de cada um dos três anotadores. Nessa análise, considerou-se equívocos instâncias em que mais de um anotador rotulou a sinopse com um gênero ao qual ela não pertencia, e como acertos instâncias em que mais de um anotador rotulou a sinopse corretamente. Acertos e equívocos foram então separados entre "parciais", quando

apenas dois dos anotadores deram a mesma resposta na rotulação, e “totais”, quando todos os anotadores concordaram com a resposta atribuída. A Tabela 4.5 apresenta todos os acertos e equívocos por gênero feitos pelos anotadores, normalizados pela frequência de cada gênero na base P-TMDb(H).

Tabela 4.5: Contagem de Acertos e Equívocos por gênero para a classificação manual da base P-TMDb(H)

	Acertos Totais	Acertos Parciais	Equívocos Totais	Equívocos Parciais
Ação	0,148	0,222	0,019	0,074
Aventura	0,256	0,308	0,154	0,205
Animação	0,217	0,435	0,000	0,000
Comédia	0,205	0,227	0,023	0,023
Crime	0,476	0,286	0,000	0,238
Documentário	0,500	0,333	0,000	0,167
Drama	0,329	0,329	0,105	0,171
Família	0,000	0,261	0,043	0,130
Fantasia	0,556	0,259	0,222	0,481
História	0,211	0,211	0,053	0,211
Horror	0,333	0,222	0,056	0,000
Música	0,308	0,308	0,000	0,154
Mistério	0,267	0,133	0,067	0,333
Romance	0,400	0,280	0,040	0,320
Ficção Científica	0,286	0,321	0,071	0,036
TV	0,000	0,000	0,000	0,286
Suspense	0,227	0,341	0,114	0,114
Guerra	0,471	0,118	0,176	0,000

Fonte: Autoria Própria

A análise de acertos e equívocos se mostrou interessante não apenas para julgar o nível de concordância entre os anotadores por gênero, mas também como indicativo de gêneros que possam ter características mais explícitas em termos da sua classificação somente pela leitura de suas sinopses.

Pode-se perceber que gêneros como Fantasia, Documentário, Crime, Guerra e Romance tiveram os mais altos níveis de acertos totais, mostrando que suas características podem ser bem definidas e facilmente identificadas por humanos. No entanto, gêneros como Família e TV não apresentaram acertos totais, mostrando que suas características principais não são bem definidas e não puderam ser identificadas com a mesma facilidade pelos humanos.

Equívocos também podem ser analisados no contexto em que anotadores tenham identificado gêneros em sinopses que não os possuíam. Tal fato pode indicar que as características desses gêneros podem ser muito genéricas e identificadas em várias sinopses,

como é o caso do gênero Fantasia, que possui a maior proporção de acertos totais, mas também as maiores proporções de equívocos totais e parciais entre todos os gêneros. Equívocos também podem indicar que um determinado gênero não possui uma boa definição para os anotadores, como é o caso de TV, para o qual não se obteve acertos, mas apenas equívocos parciais.

A análise dos resultados obtidos ao final do experimento com os anotadores humanos, embora não podendo servir como base de comparação direta para os demais resultados obtidos neste trabalho, apresentou então algumas características acerca da performance humana na classificação de gêneros cinematográficos pela análise de sua sinopse que puderam ser utilizadas para melhor guiar o restante da metodologia aplicada neste trabalho.

4.3 EXPERIMENTOS COM CLASSIFICADORES AUTOMÁTICOS

Uma das principais tarefas deste trabalho foi a realização da classificação de gêneros cinematográficos por meio de algoritmos de aprendizado de máquina. Sendo assim, esta seção apresenta a metodologia geral empregada na realização desses experimentos.

Para o treinamento e teste dos classificadores foi empregada a Validação Cruzada de 10 partições. Cada base de dados, P-TMDb, P-TMDb(-) e P-TMDb(+), foi aleatoriamente dividida em 10 partições que foram então utilizadas para criar 10 conjuntos de treinamento e teste. Uma vez que os conjuntos de treinamento e teste foram separados, a etapa de extração de características foi realizada, gerando os arquivos de treinamento e teste no formato ARFF. Em seguida os arquivos ARFF foram importados para o ambiente MEKA, que implementa os classificadores utilizados neste trabalho. Ao final da classificação, os resultados das 10 partições de cada experimento foram compilados por uma média simples para que se obtenha os resultados finais de cada experimento.

As subseções a seguir descrevem os conjuntos de características extraídos para uso neste trabalho, assim como os classificadores utilizados nos experimentos. Em seguida, os resultados dos classificadores são apresentados e analisados.

4.3.1 Características Extraídas

Para a realização dos experimentos deste trabalho, diversos grupos de extração de características textuais foram elaborados. Cada grupo descrito a seguir é composto de diversos conjuntos de características distintos agrupados por possuírem técnicas similares de extração. Práticas

como a extração de radicais (*stemming*) e remoção de palavras vazias (*stopwords*) foram utilizadas na extração de todos os conjuntos de características nas quais pudessem ser aplicadas. Adicionalmente, em casos em que etiquetadores POS (*Part Of Speech*) foram necessários para a extração de características, o conjunto de ferramentas NLTK apresentado por Bird et al. (2009) foi utilizado quando nenhum outro etiquetador é mencionado explicitamente. Os parágrafos a seguir descrevem os grupos e conjuntos de características empregados.

O primeiro grupo de características (G1) é formado por características que representam informações estruturais simples da sinopse. As características C1.1, C1.2 e C1.3 representam, respectivamente, a contagem de sentenças, de palavras e de caracteres da sinopse, incluindo-se os sinais de pontuação.

O segundo grupo de características (G2) é formado por conjuntos de característica baseados na utilização da métrica *Term Frequency - Inverse Document Frequency* (TF-IDF) dos termos da sinopse analisada. Para a utilização da TF-IDF neste trabalho, valores de TF-IDF foram extraídos para cada unigrama, bigrama e trigrama presentes no conjunto de treinamento. Em seguida o teste estatístico chi-quadrado de Pearson (1900) foi utilizado para que selecionar os 1.000 n-gramas com maior probabilidade de conterem termos característicos para a base estudada. Ao final da extração, o conjunto de características extraído para a sinopse é um vetor de 1.000 dimensões contendo os valores TF-IDF dos n-gramas selecionados para a sinopse representada.

Quatro conjuntos de características distintos foram extraídos neste grupo, utilizando-se da extração descrita acima e variando os termos considerados na extração dos n-gramas. Os conjuntos C2.1, C2.2, C2.3 e C2.4 representam respectivamente versões do conjunto nos quais todos os termos presentes na sinopse foram considerados na extração, somente termos classificados como substantivos foram considerados, somente termos adjetivos e somente termos classificados como verbos foram considerados.

O terceiro grupo de características (G3) corresponde às características que indicam a presença do nome de cada gênero na sinopse analisada. O conjunto C3.1 é composto por um vetor binário de 18 dimensões que procura explicitamente pela presença dos termos “ação”, “aventura”, “animação”, “comédia”, “crime”, “documentário”, “drama”, “família”, “fantasia”, “história”, “horror”, “música”, “mistério”, “romance”, “ficção”, “tv”, “suspense” e “guerra” no texto de cada sinopse.

O quarto grupo de características (G4) explora a construção de dicionários de termos mais frequentes para cada gênero. Tais dicionários foram extraídos, de modo geral, fazendo-se a contagem de ocorrência de cada termo na base original. Em seguida os 100 termos mais

frequentes por gênero foram compilados em dicionários individuais.

Para a extração dos conjuntos de características, cada termo da sinopse foi então checado para verificar sua presença em cada um dos dicionários gerados, sendo que em caso positivo o valor correspondente ao gênero no vetor de características é acrescido em 1. Ao final da extração, a sinopse é representada pelo número de ocorrências de seus termos nos dicionários de cada gênero.

Os conjuntos de características C4.1, C4.2, C4.3 e C4.4 foram extraídos, considerando-se os tipos de termos usados para a construção de seu dicionário, respectivamente todos os termos foram considerados, somente termos classificados como substantivos foram considerados, somente termos classificados como adjetivos foram considerados, e somente termos classificados como verbos foram considerados.

O quinto grupo de características (G5) é composto por características que fazem uso da biblioteca NLPNET (*Natural Language Processing with neural networks*). A biblioteca, apresentada por Fonseca e Rosa (2013), é desenvolvida na linguagem *Python* e realiza diversas tarefas de processamento de linguagens naturais utilizando-se de redes neurais. Neste trabalho, a NLPNET foi utilizada pelo seu etiquetador POS (*Part-Of-Speech*) que tem a função de atribuição de classes gramaticais da língua portuguesa aos termos de uma sentença.

As características C5.1, C5.2, C5.3, C5.4 e C5.5 extraídas para este grupo representam a contagem de termos pertencentes a determinadas classes gramaticais na sinopse analisada. As características representam, respectivamente, a contagem de verbos, substantivos, adjetivos, advérbios e pronomes da sinopse.

O sexto grupo de características (G6) codifica aspectos linguísticos das sinopses. As características foram apresentadas por Zhou et al. (2004) no contexto da detecção de *deception*, mas mesmo que tais características tenham sido previamente utilizadas para capturar informações linguísticas em outro contexto, elas foram escolhidas para análise neste estudo por conterem informações sobre aspectos linguísticos que podem caracterizar a escrita de determinados gêneros cinematográficos que possam não ser detectadas por outros grupos de característica estudados.

As características de aspectos linguísticos são: *pausality*, definida como a contagem de sinais de pontuação da sinopse em relação ao seu número de sentenças; *emotiveness*, definida pelo número total de adjetivos e advérbios em relação ao número total de verbos e substantivos; *uncertainty*, definida pelo total de verbos modais e usos da voz passiva contidos na sinopse; e *non immediacy*, definida pela contagem de pronomes na primeira e segunda pessoa.

As características C6.1, C6.2, C6.3 e C6.4 representam as métricas *pausality*, *emotiveness*, *uncertainty* e *non immediacy*, respectivamente.

O sétimo grupo de características (G7) utiliza o dicionário LIWC (*Linguistic Inquiry and Word Count*) para calcular a frequência relativa de emoções, estilos de pensamento, considerações sociais, e etiquetas POS na sinopse analisada. Neste trabalho foi empregada a versão em português do LIWC, apresentada por Filho et al. (2013) e baseada no dicionário LIWC2001 (TAUSCZIK; PENNEBAKER, 2010).

Segundo os autores, o dicionário LIWC é composto por cerca de 27.149 termos anotados com um total de 64 classes semânticas que classificam cada termo em relação a aspectos como sua classe gramatical, emoção representada, relação social e familiar, aspecto cognitivo apresentado, processo perceptivo apresentado, processo biológico apresentado, principais impulsos e necessidades, tempo verbal, relatividade, preocupações pessoais, palavreado informal e pontuações. O conjunto de características C7.1 é representado por um vetor de 64 dimensões contendo uma contagem normalizada para cada classe semântica presente no dicionário.

O oitavo grupo de características (G8) é composto por características extraídas pelo uso do modelo *Latent Dirichlet Allocation* (LDA) (BLEI et al., 2003) para a classificação da sinopse analisada em 50 tópicos. A implementação do algoritmo de extração do modelo LDA utilizada neste trabalho é disponibilizada pelo *toolkit* Gensim (REHUREK; SOJKA, 2010) e é baseada no algoritmo apresentado por Hoffman et al. (2010).

O conjunto de características C8.1 é representado então por um vetor de 50 dimensões no qual cada posição contém um valor entre 0 e 1 representando a probabilidade da sinopse analisada pertencer a um determinado tópico.

O nono grupo de características (G9) é baseado no uso de *embeddings*, os quais, segundo Hartmann et al. (2017), são vetores de números reais capazes de representar termos e suas relações em espaços n-dimensionais de forma que informações sintáticas, semânticas e morfológicas sobre esses termos sejam capturadas. Os autores notam que diversas abordagens para a codificação de termos em *embeddings* existem na literatura, como a abordagem preditiva apresentada por Mikolov et al. (2013), na qual o modelo é treinado para reconhecer o contexto em que cada termo está inserido, com o objetivo de prever termos vizinhos a partir de um ou mais termos presentes em seu contexto. Essa abordagem é comumente implementada com o nome *Word2Vec*.

Como o *Word2Vec* é utilizado para a geração de *embeddings* que representam termos,

outro modelo precisa ser utilizado para que se represente documentos. Esse é o problema abordado por Le e Mikolov (2014) em seu trabalho que descreve o *Doc2Vec*, no qual o *Word2Vec* é alterado de forma que, no contexto de cada termo, seja inserido um identificador do parágrafo único ao qual o termo pertença, permitindo assim que informações sobre o tópico do parágrafo possam ser integradas ao modelo.

Neste trabalho, uma vez que os vetores de parágrafos foram obtidos para os conjuntos de treinamento, o vetor de características de cada instância foi obtido pela obtenção da posição de cada termo contido nela no espaço vetorial construído. Em seguida a média ponderada de todos os vetores da sinopse foi obtida utilizando-se dos valores TF-IDF de cada termo, resultando no vetor final que representa cada sinopse estudada.

Os conjuntos de características C9.1, C9.2 e C9.3 representam intâncias do modelo *Doc2Vec* nas quais o mesmo foi construído e utilizado para a extração de características de vetores de 50, 100 e 1000 dimensões respectivamente. Já o conjunto C9.4 faz uso do modelo pré-treinado *Wang2Vec* disponibilizado por Hartmann et al. (2017) com vetores de 50 dimensões.

O quadro 4.1 compila a lista dos grupos de características extraídas neste trabalho, incluindo os conjuntos de características pertencentes a cada grupo e breves descrições sobre eles. Uma lista completa das características utilizadas neste trabalho, assim como suas descrições e dimensionalidades pode ser encontrada no Quadro 4.2.

Quadro 4.1: Resumo dos grupos de características utilizados no trabalho

Grupo	Conjuntos de Características	Descrição
G1	C1.1, C1.2, C1.3	Características estruturais simples
G2	C2.1, C2.2, C2.3, C2.4	Valores TF-IDF
G3	C3.1	Presença do nome de cada gênero
G4	C4.1, C4.2, C4.3, C4.4	Construção de dicionários de termos mais frequentes por gênero
G5	C5.1, C5.2, C5.3, C5.4, C5.5	Contagem de classes gramaticais
G6	C6.1, C6.2, C6.3, C6.4	Outros aspectos linguísticos
G7	C7.1	Contagem normalizada de classes do LIWC
G8	C8.1	Uso do modelo LDA
G9	C9.1, C9.2, C9.3, C9.4	Uso de modelos baseados em *embeddings*

Fonte: Autoria Própria

Quadro 4.2: Lista completa das características utilizadas neste trabalho

Grupo	Característica	Descrição	Dimensionalidade
G1	C1.1	Contagem de sentenças da sinopse	1
G1	C1.2	Contagem de palavras da sinopse	1
G1	C1.3	Contagem de caracteres da sinopse	1
G2	C2.1	TF-IDF considerando todos os termos	1000
G2	C2.2	TF-IDF considerando apenas substantivos	1000
G2	C2.3	TF-IDF considerando apenas adjetivos	1000
G2	C2.4	TF-IDF considerando apenas verbos	1000
G3	C3.1	Presença do nome de cada gênero	18
G4	C4.1	Dicionários de todos os termos mais frequentes por gênero	100
G4	C4.2	Dicionários de substantivos mais frequentes por gênero	100
G4	C4.3	Dicionários de adjetivos mais frequentes por gênero	100
G4	C4.4	Dicionários de verbos mais frequentes por gênero	100
G5	C5.1	Contagem de verbos da sinopse	1
G5	C5.2	Contagem de substantivos da sinopse	1
G5	C5.3	Contagem de adjetivos da sinopse	1
G5	C5.4	Contagem de advérbios da sinopse	1
G5	C5.5	Contagem de pronomes da sinopse	1
G6	C6.1	Aspecto linguístico *pausality*	1
G6	C6.2	Aspecto linguístico *emotiveness*	1
G6	C6.3	Aspecto linguístico *uncertainty*	1
G6	C6.4	Aspecto linguístico *non immediacy*	1
G7	C7.1	Uso do dicionário LIWC	64
G8	C8.1	Uso do modelo LDA	50
G9	C9.1	Uso do modelo *Doc2Vec*	50
G9	C9.2	Uso do modelo *Doc2Vec*	100
G9	C9.3	Uso do modelo *Doc2Vec*	1000
G9	C9.4	Uso do modelo *Wang2Vec*	50

Fonte: Autoria Própria

4.3.2 Experimentos Elaborados

Para a realização deste trabalho, experimentos de avaliação foram planejados visando que as características descritas acima fossem avaliadas: individualmente; com possíveis combinações entre conjuntos de características do mesmo grupo; e com possíveis combinações entre características de grupos distintos. Neste último caso, foram exploradas as abordagens de fusão imediata (*early fusion*) e de fusão tardia (*late fusion*).

Os experimentos iniciais foram compostos em sua maioria por instâncias nas quais cada característica foi testada individualmente. Adicionalmente, experimentos contendo a combinação de todos as características de um mesmo grupo de característica foram elaboradas quando possível.

Todos os Experimentos Individuais (EI) foram realizados com bases P-TMDb, P-TMDb(+) e P-TMDb(-). A relação entre os 31 experimentos individuais e dos grupos de

características que foram utilizadas em cada experimento pode ser observada no Quadro 4.3.

Quadro 4.3: Relação dos Experimentos Individuais conduzidos

Experimento	Conjuntos de Características	Experimento	Conjuntos de Características
EI_00	C3.1	EI_17	C5.5
EI_01	C1.1	EI_18	C5.1, C5.2, C5.3, C5.4, C5.5
EI_02	C1.2		
EI_03	C1.3	EI_19	C6.1
EI_04	C1.1, C1.2, C1.3	EI_20	C6.2
EI_05	C2.1	EI_21	C6.3
EI_06	C2.2	EI_22	C6.4
EI_07	C2.3	EI_23	C6.1, C6.2, C6.3, C6.4
EI_08	C2.4		
EI_09	C4.1	EI_24	C7.1
EI_10	C4.2	EI_25	C8.1
EI_11	C4.3	EI_26	C5.1, C5.2, C5.3, C5.4, C5.5, C6.1, C6.2, C6.3, C6.4
EI_12	C4.4		
EI_13	C5.1	EI_27	C9.1
EI_14	C5.2	EI_28	C9.2
EI_15	C5.3	EI_29	C9.3
EI_16	C5.4	EI_30	C9.4

Fonte: Autoria Própria

Os demais experimentos conduzidos neste estudo foram elaborados para testar uma lista semi-exaustiva de combinações entre os grupos de características. Para tal, combinações entre os conjuntos de características de cada grupo foram escolhidos para representar o grupo como um todo.

Os conjuntos de características escolhidos para essa etapa foram os que foram julgados como contendo o maior número de informações que o grupo poderia oferecer, isto é, não continham restrições acerca das classes gramaticais dos termos utilizados em sua extração, continham todas as características de um mesmo grupo quando possível, ou apresentavam a variação da característica com vetor de maior dimensionalidade entre variações similares do mesmo grupo. Esses conjuntos foram denominados Conjuntos Máximos (CM) e podem ser observados no Quadro 4.4.

Após a definição do Conjunto Máximo para cada grupo de características, um total de 27 Experimentos Combinatoriais (EC) foram conduzidos com o intuito de explorar de forma semi-exaustiva as possíveis combinações dos conjuntos máximos.

Devido às limitações de tempo e recursos, um estudo verdadeiramente exaustivo de combinações de conjuntos máximos não poderia ser realizado. Sendo assim, algumas diretrizes foram empregadas no planejamento dos experimentos combinatoriais, a saber: (i) os conjuntos máximos CM2, CM4 e CM9 são mutualmente exclusivos devido ao tamanho de seus

Quadro 4.4: Relação das características que compõem cada Conjunto Máximo

Conjuntos Máximos	Conjuntos de Características
CM1	C1.1, C1.2, C1.3
CM2	C2.1
CM3	C3.1
CM4	C4.1
CM5	C5.1, C5.2, C5.3, C5.4, C5.5
CM6	C6.1, C6.2, C6.3, C6.4
CM7	C7.1
CM8	C8.1
CM9	C9.3

Fonte: Autoria Própria

vetores de características ou por possuírem métodos de extração que resultam em características similares; (ii) os conjuntos máximos CM1 e CM3 foram testados sempre em combinação por possuírem vetores de características pequenos; (iii) todas as combinações deveriam conter um dos conjuntos máximos CM2, CM4 ou CM9. Dentre as combinações possíveis é então dada a prioridade à combinações que empregam todos os outros conjuntos máximos e os conjuntos CM2, CM4 e CM9 individualmente, assim como combinações em que todos os conjuntos restantes estejam combinados com os mesmos três conjuntos máximos ao mesmo tempo. Combinações adicionais poderiam então ser testadas de acordo com a disponibilidade de recursos deste trabalho.

Uma vez que as limitações dos Experimentos Combinatoriais foram definidas, foi feito o planejamento dos experimentos a serem realizados. O Quadro 4.5 apresenta a lista dos Experimentos Combinatoriais conduzidos e os Conjuntos Máximos que os compõem, sendo que na realização desses experimentos foram testadas tanto a estratégia de fusão imediata quanto de fusão tardia.

4.3.3 Classificadores Utilizados

Quatro algoritmos de classificação multirrótulo foram utilizados neste trabalho. Esses algoritmos foram escolhidos de forma a explorar abordagens de classificação distintas e por terem implementações disponíveis no *tookit* de aprendizado multirrótulo MEKA (READ et al., 2016).

O primeiro classificador utiliza o algoritmo *Binary Relevance* (Relevância Binária) descrito na Subseção 2.4.1. Segundo a documentação do *toolkit* MEKA, o algoritmo de

Quadro 4.5: Relação dos Experimentos Combinatoriais conduzidos

Experimento Combinatorial	Conjuntos Máximos
EC_1	CM1, CM3 , CM4
EC_2	CM4, CM5
EC_3	CM4, CM6
EC_4	CM4, CM7
EC_5	CM4, CM8
EC_6	CM1, CM3, CM4, CM5
EC_7	CM4, CM5, CM6
EC_8	CM4, CM7, CM8
EC_9	CM1, CM3, CM4, CM5, CM6, CM7, CM8
EC_10	CM1, CM3, CM9
EC_11	CM5, CM9
EC_12	CM6, CM9
EC_13	CM7, CM9
EC_14	CM8, CM9

Experimento Combinatorial	Conjuntos Máximos
EC_15	CM1, CM3, CM5, CM9
EC_16	CM5, CM6, CM9
EC_17	CM7, CM8, CM9
EC_18	CM1, CM3, CM5, CM6, CM7, CM8, CM9
EC_19	CM1, CM2, CM3
EC_20	CM2, CM5
EC_21	CM2, CM6
EC_22	CM2, CM7
EC_23	CM2, CM8
EC_24	CM1, CM2, CM3, CM5
EC_25	CM2, CM5, CM6
EC_26	CM2, CM7, CM8
EC_27	CM1, CM2, CM3, CM5, CM6, CM7, CM8

Fonte: Autoria Própria

classificação por relevância binária utilizado neste trabalho foi baseado na implementação encontrada no *framework* MULAN, apresentado por Tsoumakas et al. (2009). O segundo classificador utilizada a abordagem *Classifier Chains* (Correntes de Classificadores), conforme descrita na Subseção 2.4.2. O terceiro classificador utiliza o algoritmo RAkEL (*Random k-labelsets*) descrito na Subseção 2.4.3. Neste estudo, o classificador RAkEL foi inicializado para trabalhar com 10 subconjuntos distintos de 3 rótulos cada. O último classificador multirrótulo utilizado neste estudo fez uso de DBPNN (*Deep Back-Propagation Neural Network*), a qual tem sua inicialização de pesos usando-se de RBMs (*Restricted Boltzmann Machines*), como descrito na Subseção 2.4.4. O classificador DBPNN utilizado foi instanciado utilizando duas RBMs, 10 unidades ocultas, taxa de aprendizado $0,1$ e *momentum* $0,1$.

Todos os classificadores multirrótulo descritos anteriormente foram instanciados utilizando um classificador base multiclasse. O classificador base escolhido para esse fim foi o J48, o qual é uma implementação livre do algoritmo C4.5 , originalmente apresentado por Quinlan (2014), disponibilizada pelo ambiente WEKA (WITTEN et al., 2016). Esse é um algoritmo baseado em árvores de decisão, nas quais as características são escolhidas para compor os nós da árvore de acordo com seus valores de ganho de informação. O algoritmo J48 utilizado neste estudo foi inicializado com um limiar de poda de $0,25$ e com um número mínimo de 2 instâncias por folha.

Cabe destacar que um classificador base SVM (*Support Vector Machines*) também foi

considerado para estudo, mas devido às limitações de tempo e ao fato dos experimentos iniciais utilizando tal classificador não terem mostrado resultados superiores àqueles do J48, o uso de tal classificador foi reservado para possíveis trabalhos futuros.

5

RESULTADOS E DISCUSSÕES

Nesta seção são apresentados e discutidos os resultados obtidos nos Experimentos Individuais e Combinatoriais com as bases P-TMDb, P-TMDb(-) e P-TMDb(+) usando os classificadores descritos na seção anterior. Todos os resultados apresentados nesta seção representam a média das métricas precisão, revocação e medida-F obtidas por validação cruzada de 10 partições, sendo que em geral o desvio padrão de cada média não ultrapassou $0,015$.

5.1 EXPERIMENTOS INDIVIDUAIS

Para análise dos resultados dos experimentos individuais, primeiramente são apresentadas as métricas obtidas com a base P-TMDb. A Tabela 5.1 apresenta os resultados completos obtidos nesse caso, enquanto a Figura 5.1 apresenta uma visualização comparativa das medidas-F obtidas em cada experimento.

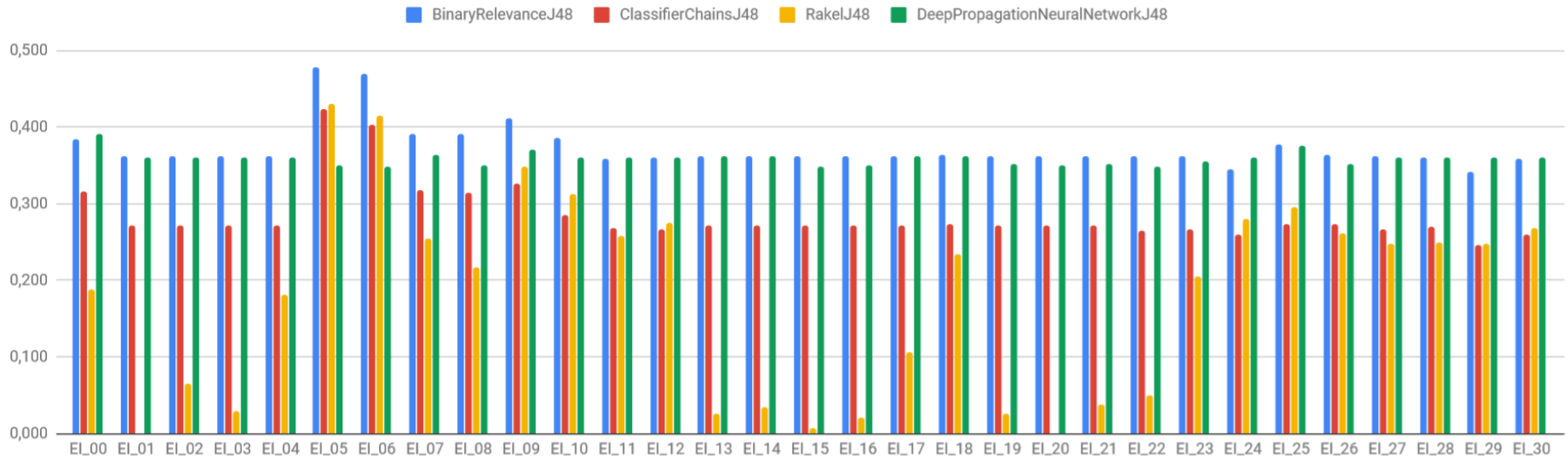
Na Tabela 5.1 e na Figura 5.1 é possível observar que o classificador *BinaryRelevanceJ48* obteve os melhores resultados por possuir, em geral, resultados equivalentes ou superiores aos outros classificadores. Nota-se que o melhor resultado, em geral, é obtido com o experimento EI_05, que utiliza o conjunto de características baseadas em TF-IDF com todos os termos da base considerados na extração, obtendo uma medida-F de $0,478$, precisão de $0,477$ e revocação de $0,480$. Tal resultado é condizente com os apresentados por outros autores para o domínio da classificação multirrótulo de gêneros cinematográficos utilizando características textuais de suas sinopses (Tabela 3.2) e ligeiramente inferior aos apresentados pela anotação manual (Tabela 4.4).

Tabela 5.1: Resultados dos Experimentos Individuais na base P-TMDb

Exp.	BinaryRelevanceJ48			ClassifierChainsJ48			RakelJ48			DeepPropagationNeuralNetworkJ48		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
EI_00	0,330	0,462	0,385	0,485	0,235	0,316	0,572	0,113	0,189	0,386	0,397	0,392
EI_01	0,318	0,420	0,362	0,444	0,196	0,272	0,000	0,000	0,000	0,317	0,419	0,361
EI_02	0,318	0,420	0,362	0,444	0,196	0,272	0,458	0,035	0,065	0,317	0,419	0,361
EI_03	0,318	0,420	0,362	0,444	0,196	0,272	0,401	0,016	0,030	0,317	0,419	0,361
EI_04	0,318	0,420	0,362	0,444	0,196	0,271	0,400	0,118	0,182	0,317	0,419	0,361
EI_05	0,477	0,480	0,478	0,505	0,365	0,424	0,453	0,411	0,431	0,350	0,350	0,350
EI_06	0,453	0,489	0,469	0,500	0,337	0,402	0,468	0,372	0,415	0,349	0,349	0,349
EI_07	0,388	0,393	0,390	0,464	0,241	0,317	0,438	0,180	0,255	0,326	0,413	0,363
EI_08	0,384	0,398	0,391	0,457	0,239	0,314	0,416	0,148	0,218	0,350	0,350	0,350
EI_09	0,410	0,412	0,411	0,436	0,261	0,326	0,314	0,390	0,348	0,354	0,390	0,371
EI_10	0,381	0,393	0,387	0,408	0,219	0,285	0,280	0,353	0,312	0,317	0,419	0,361
EI_11	0,330	0,395	0,359	0,435	0,194	0,269	0,232	0,289	0,257	0,317	0,419	0,361
EI_12	0,336	0,392	0,361	0,429	0,194	0,267	0,248	0,308	0,275	0,317	0,419	0,361
EI_13	0,318	0,420	0,362	0,444	0,196	0,272	0,414	0,014	0,027	0,318	0,419	0,361
EI_14	0,318	0,420	0,362	0,443	0,196	0,272	0,346	0,019	0,034	0,320	0,418	0,362
EI_15	0,318	0,420	0,362	0,444	0,196	0,272	0,142	0,004	0,007	0,348	0,349	0,348
EI_16	0,318	0,420	0,362	0,444	0,196	0,272	0,259	0,011	0,020	0,351	0,351	0,351
EI_17	0,318	0,420	0,362	0,443	0,195	0,271	0,487	0,060	0,107	0,319	0,417	0,361
EI_18	0,318	0,422	0,363	0,445	0,197	0,273	0,318	0,186	0,235	0,319	0,418	0,362
EI_19	0,318	0,420	0,362	0,444	0,196	0,272	0,501	0,013	0,025	0,334	0,373	0,352
EI_20	0,318	0,420	0,362	0,444	0,196	0,272	0,000	0,000	0,000	0,349	0,350	0,349
EI_21	0,318	0,420	0,362	0,444	0,196	0,272	0,391	0,020	0,037	0,352	0,352	0,352
EI_22	0,318	0,420	0,362	0,434	0,191	0,266	0,504	0,027	0,051	0,348	0,349	0,349
EI_23	0,318	0,420	0,362	0,435	0,193	0,267	0,363	0,143	0,205	0,329	0,389	0,355
EI_24	0,342	0,347	0,345	0,307	0,225	0,260	0,236	0,346	0,281	0,317	0,419	0,361
EI_25	0,369	0,386	0,377	0,381	0,214	0,274	0,275	0,320	0,296	0,364	0,387	0,375
EI_26	0,320	0,421	0,364	0,441	0,198	0,274	0,268	0,254	0,261	0,350	0,353	0,352
EI_27	0,318	0,422	0,362	0,432	0,193	0,267	0,223	0,277	0,247	0,317	0,419	0,361
EI_28	0,321	0,412	0,361	0,432	0,196	0,269	0,217	0,291	0,249	0,317	0,419	0,361
EI_29	0,331	0,353	0,341	0,345	0,191	0,246	0,206	0,310	0,247	0,317	0,419	0,361
EI_30	0,337	0,387	0,359	0,388	0,196	0,260	0,224	0,331	0,268	0,317	0,419	0,361

Fonte: Autoria Própria

Figura 5.1: Medida-F de cada classificador para cada Experimento Individual na base P-TMDb



Fonte: Autoria Própria

Os classificadores *BinaryRelevanceJ48* e *DeepPropagationNeuralNetworkJ48* obtiveram resultados similares, geralmente não apresentando valores de medida-F inferiores a $0,362$ mesmo ao se usar conjuntos de características que não capturam grande quantidade de informações sobre a sinopse. Tal fato pode indicar que parte da qualidade de predição desses classificadores se dê pelo aprendizado apenas da frequência de cada classe na base estudada, sendo que somente nos experimentos EI_01, EI_05, EI_06, EI_07, EI_08, EI_09, EI_010, EI_024, EI_025 e EI_029, o classificador *BinaryRelevanceJ48* conseguiu apresentar resultados significativamente distintos em relação a sua métrica medida-F. Pode-se notar que tais experimentos implementam todas as características pertencentes aos grupos de características C2, C3, C7 e C8, sendo que também apresentam características dos grupos C4 e C9.

O classificador *ClassifierChainsJ48* apresentou a mesma tendência de classificação observada para o classificador *BinaryRelevanceJ48*, sendo que seu valor mínimo de medida-F foi próximo a $0,272$. Já o classificador *RakelJ48* apresentou uma grande variação de valores resultantes, aparentando não estabelecer um valor mínimo de classificação como os outros e, ao mesmo tempo, apresentado os segundos melhores resultados dos experimentos para E_05 e E_06.

Uma vez definidos os resultados para os Experimentos Individuais na base P-TMDb, pode-se analisar os resultados dos demais experimentos para que se possa compreender o impacto do uso de outras técnicas de classificação e derivações da base de dados original. A Tabela 5.2 mostra os valores de precisão, revocação e medida-F para os Experimentos Individuais quando realizados com base P-TMDb(-). A Figura 5.2 apresenta uma visualização comparativa dos valores medida-F de cada um dos experimentos descritos acima.

Na Tabela 5.2 e na Figura 5.2 é possível observar que o rebalanceamento da base P-TMDb não teve grande impacto em relação aos valores de medida-F dos classificadores *BinaryRelevanceJ48*, *DeepPropagationNeuralNetworkJ48* e *ClassifierChainsJ48*, fazendo com que seus valores mínimos apresentassem pouca variação em relação aos obtidos na execução dos experimentos com a base original. Os classificadores continuaram tendo melhor resultado no experimento E_05, desta vez apresentando um valor de $0,475$.

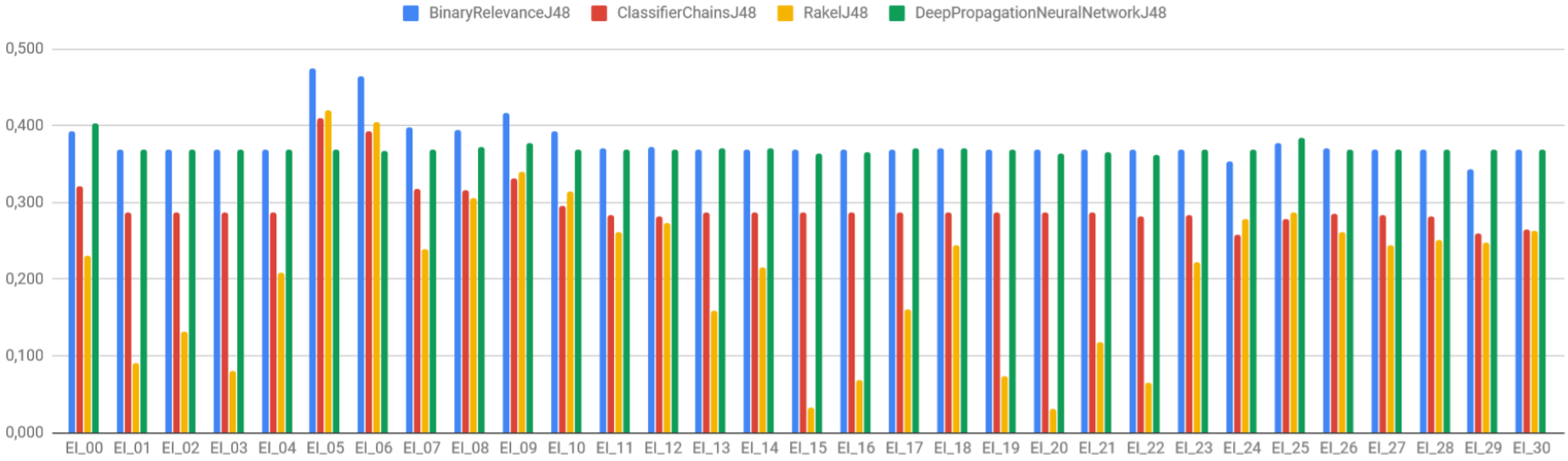
O maior impacto do algoritmo de rebalanceamento pode ser observado para o classificador *RakelJ48*, para o qual experimentos como E_01, E_02 e E_03 apresentam diferenças consideráveis em relação aos resultados obtidos na base original. No entanto, tal fato pode ser atribuído tanto ao rebalanceamento da base quanto à natureza do classificador, que se utiliza de aleatoriedade na divisão dos conjuntos de rótulos utilizados em seu treinamento,

Tabela 5.2: Resultados dos Experimentos Individuais na base P-TMDb(-)

Exp.	BinaryRelevanceJ48			ClassifierChainsJ48			RakelJ48			DeepPropagationNeuralNetworkJ48		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
EI_00	0,336	0,471	0,392	0,501	0,236	0,321	0,562	0,145	0,230	0,400	0,406	0,403
EI_01	0,322	0,434	0,369	0,463	0,208	0,287	0,491	0,051	0,091	0,322	0,434	0,369
EI_02	0,322	0,434	0,369	0,463	0,208	0,287	0,469	0,078	0,132	0,322	0,434	0,369
EI_03	0,322	0,434	0,369	0,463	0,208	0,287	0,469	0,044	0,080	0,322	0,434	0,369
EI_04	0,322	0,434	0,369	0,463	0,208	0,287	0,416	0,140	0,209	0,322	0,434	0,369
EI_05	0,474	0,476	0,475	0,496	0,348	0,409	0,452	0,391	0,419	0,369	0,369	0,369
EI_06	0,451	0,479	0,464	0,501	0,324	0,393	0,466	0,357	0,404	0,368	0,368	0,368
EI_07	0,394	0,400	0,398	0,466	0,241	0,317	0,440	0,165	0,240	0,369	0,369	0,369
EI_08	0,393	0,396	0,395	0,467	0,239	0,316	0,449	0,231	0,305	0,372	0,372	0,372
EI_09	0,415	0,420	0,417	0,450	0,262	0,331	0,311	0,375	0,340	0,372	0,382	0,377
EI_10	0,381	0,408	0,394	0,441	0,223	0,296	0,287	0,348	0,314	0,322	0,434	0,369
EI_11	0,329	0,426	0,370	0,457	0,205	0,283	0,238	0,289	0,261	0,322	0,434	0,370
EI_12	0,339	0,417	0,372	0,449	0,205	0,281	0,254	0,298	0,274	0,322	0,433	0,369
EI_13	0,322	0,434	0,369	0,463	0,208	0,287	0,426	0,100	0,159	0,323	0,433	0,370
EI_14	0,322	0,434	0,369	0,463	0,208	0,287	0,491	0,138	0,216	0,330	0,427	0,371
EI_15	0,322	0,434	0,369	0,463	0,208	0,287	0,343	0,017	0,033	0,364	0,364	0,364
EI_16	0,322	0,434	0,369	0,463	0,208	0,287	0,422	0,038	0,069	0,365	0,365	0,365
EI_17	0,322	0,434	0,369	0,463	0,208	0,287	0,512	0,096	0,161	0,329	0,425	0,371
EI_18	0,322	0,436	0,370	0,459	0,209	0,287	0,355	0,186	0,244	0,325	0,433	0,371
EI_19	0,322	0,434	0,369	0,463	0,208	0,287	0,478	0,041	0,075	0,322	0,434	0,369
EI_20	0,322	0,434	0,369	0,463	0,208	0,287	0,479	0,016	0,031	0,364	0,365	0,364
EI_21	0,322	0,434	0,369	0,463	0,208	0,287	0,487	0,068	0,118	0,366	0,367	0,366
EI_22	0,322	0,434	0,369	0,454	0,204	0,281	0,516	0,035	0,065	0,363	0,363	0,363
EI_23	0,322	0,434	0,369	0,457	0,206	0,283	0,375	0,158	0,222	0,322	0,434	0,369
EI_24	0,350	0,355	0,353	0,317	0,218	0,258	0,237	0,341	0,279	0,322	0,434	0,369
EI_25	0,358	0,400	0,377	0,408	0,211	0,278	0,281	0,293	0,287	0,383	0,385	0,384
EI_26	0,326	0,430	0,371	0,457	0,208	0,286	0,280	0,246	0,262	0,322	0,434	0,369
EI_27	0,322	0,435	0,370	0,455	0,206	0,284	0,229	0,263	0,245	0,322	0,434	0,369
EI_28	0,322	0,434	0,369	0,449	0,205	0,282	0,223	0,287	0,251	0,322	0,434	0,369
EI_29	0,330	0,357	0,343	0,372	0,199	0,259	0,210	0,305	0,249	0,322	0,434	0,369
EI_30	0,357	0,383	0,369	0,391	0,200	0,265	0,223	0,321	0,264	0,322	0,434	0,369

Fonte: Autoria Própria

Figura 5.2: Medida-F de cada classificador para cada Experimento Individual na base P-TMDb(-)



Fonte: Autoria Própria

requerendo uma quantidade maior de testes para que se possa fazer um julgamento mais preciso sobre seu impacto quando usado em conjunto com a base de dados rebalanceada.

A Tabela 5.3 apresenta todos os resultados obtidos da execução dos Experimentos Individuais na base P-TMDb(+).A Figura 5.3 apresenta a comparação entre os valores de medida-F obtidos nesses experimentos.

Observando a Tabela 5.3 e a Figura 5.3 é possível notar o impacto que o rebalanceamento da base P-TMDb pelo algoritmo LP-ROS teve nas classificações. Quando se tratou do melhor classificador, *BinaryRelevanceJ48*, pode-se notar que em todos os experimentos nos quais o uso do classificador apresentava resultados superiores ao valor base de sua classificação, o uso da base rebalanceada apresentou um claro impacto em relação a sua medida-F.

Os melhores resultados de classificação com a base P-TMDb(+), uma medida-F de $0,577$, foram obtidos no experimento EI_09, que fez uso de características baseadas no dicionário com os termos mais frequentes para cada gênero, considerando-se todos os termos para a construção de tal dicionário. Entre os melhores resultados também estão os do experimento EI_10, que também utilizou características baseadas no dicionário de termos mais frequentes por gênero, porém considerando apenas os termos substantivos das sinopses, os do experimento E_24, que utilizou a contagem de classes semânticas do LIWC, e os do experimento E_05, que apresentava o melhor resultado anteriormente e que faz uso de características TF-IDF.

Os resultados para o classificador *DeepPropagationNeuralNetworkJ48* não apresentaram grandes variações após o rebalanceamento da base original, enquanto que os resultados do classificador *ClassifierChainsJ48* apresentaram a mesma tendência do classificador *BinaryRelevanceJ48*. Os valores mínimos de medida-F do classificador *ClassifierChainsJ48* mantiveram-se similares, enquanto que os mesmos conjuntos de características que apresentaram mudanças positivas para o melhor classificador também obtiveram bons resultados neste. O mesmo pode ser observado para o classificador *RakelJ48*, no qual podemos mais uma vez observar que os mesmos experimentos que apresentaram mudança significativa em outros classificadores obtiveram alterações similares.

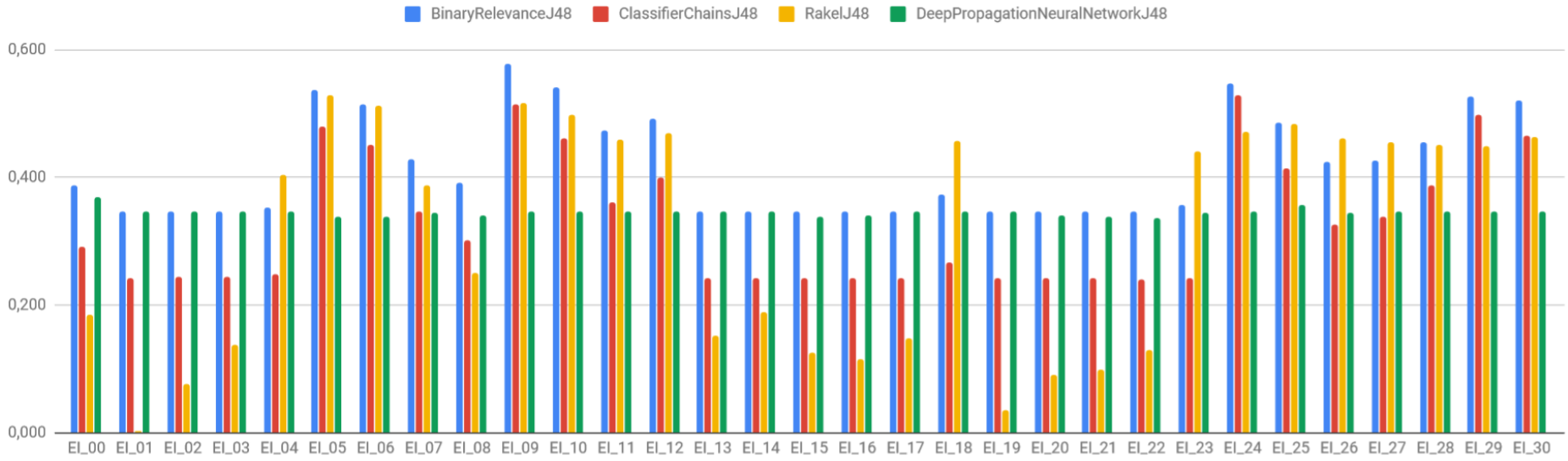
Ao fim da análise dos Experimentos Individuais foi possível perceber que o classificador *BinaryRelevanceJ48* obteve os melhores resultados em relação à medida-F para os experimentos realizados. Também pôde-se notar que a base P-TMDb(-) derivada por rebalanceamento não produziu melhora significativa para os resultados dos experimentos, enquanto que a base P-TMDb(+), que aplicou um rebalanceamento similar ao aumentar o

Tabela 5.3: Resultados dos Experimentos Individuais na base P-TMDb(+)

Exp.	BinaryRelevanceJ48			ClassifierChainsJ48			RakelJ48			DeepPropagationNeuralNetworkJ48		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
EI_00	0,350	0,436	0,388	0,483	0,208	0,290	0,575	0,110	0,184	0,355	0,387	0,370
EI_01	0,324	0,372	0,346	0,438	0,168	0,242	0,040	0,002	0,003	0,324	0,372	0,347
EI_02	0,324	0,372	0,346	0,435	0,169	0,244	0,469	0,041	0,075	0,324	0,372	0,347
EI_03	0,324	0,372	0,346	0,437	0,169	0,244	0,453	0,081	0,138	0,324	0,372	0,347
EI_04	0,327	0,384	0,353	0,434	0,173	0,248	0,484	0,346	0,403	0,324	0,372	0,347
EI_05	0,537	0,538	0,537	0,566	0,417	0,480	0,578	0,488	0,529	0,335	0,342	0,339
EI_06	0,514	0,515	0,515	0,544	0,385	0,451	0,564	0,469	0,512	0,334	0,344	0,339
EI_07	0,428	0,430	0,428	0,493	0,268	0,347	0,572	0,294	0,388	0,328	0,365	0,345
EI_08	0,349	0,446	0,391	0,460	0,224	0,301	0,581	0,160	0,250	0,328	0,355	0,341
EI_09	0,577	0,577	0,577	0,553	0,482	0,515	0,469	0,574	0,516	0,324	0,372	0,347
EI_10	0,541	0,543	0,542	0,519	0,417	0,462	0,451	0,554	0,497	0,324	0,372	0,347
EI_11	0,465	0,482	0,474	0,470	0,294	0,361	0,413	0,518	0,459	0,324	0,372	0,347
EI_12	0,490	0,493	0,492	0,487	0,340	0,400	0,424	0,524	0,469	0,324	0,372	0,347
EI_13	0,324	0,372	0,346	0,434	0,168	0,242	0,470	0,091	0,153	0,325	0,372	0,347
EI_14	0,324	0,372	0,346	0,435	0,168	0,243	0,477	0,118	0,190	0,326	0,370	0,346
EI_15	0,324	0,372	0,346	0,438	0,168	0,242	0,463	0,073	0,126	0,337	0,337	0,337
EI_16	0,324	0,372	0,346	0,437	0,168	0,242	0,463	0,066	0,116	0,341	0,341	0,341
EI_17	0,324	0,372	0,346	0,438	0,168	0,242	0,478	0,088	0,148	0,325	0,370	0,346
EI_18	0,355	0,396	0,373	0,436	0,192	0,267	0,442	0,471	0,456	0,326	0,371	0,347
EI_19	0,324	0,372	0,346	0,438	0,168	0,242	0,462	0,018	0,034	0,327	0,366	0,345
EI_20	0,324	0,372	0,346	0,438	0,168	0,242	0,466	0,050	0,091	0,336	0,347	0,341
EI_21	0,324	0,372	0,346	0,437	0,167	0,242	0,468	0,056	0,099	0,337	0,338	0,337
EI_22	0,324	0,372	0,346	0,432	0,166	0,239	0,478	0,075	0,129	0,337	0,337	0,337
EI_23	0,335	0,381	0,356	0,424	0,170	0,243	0,442	0,438	0,440	0,327	0,366	0,345
EI_24	0,546	0,549	0,548	0,526	0,531	0,528	0,415	0,548	0,472	0,324	0,372	0,347
EI_25	0,478	0,493	0,485	0,488	0,359	0,414	0,452	0,520	0,483	0,356	0,358	0,357
EI_26	0,413	0,435	0,423	0,456	0,256	0,327	0,425	0,504	0,461	0,327	0,364	0,345
EI_27	0,416	0,438	0,427	0,457	0,270	0,339	0,406	0,515	0,454	0,324	0,372	0,347
EI_28	0,445	0,468	0,456	0,474	0,328	0,387	0,400	0,518	0,451	0,324	0,372	0,347
EI_29	0,524	0,531	0,527	0,500	0,494	0,497	0,389	0,530	0,449	0,324	0,372	0,347
EI_30	0,518	0,522	0,520	0,497	0,438	0,465	0,406	0,539	0,463	0,324	0,372	0,347

Fonte: Autoria Própria

Figura 5.3: Medida-F de cada classificador para cada Experimento Individual na base P-TMDb(+)



Fonte: Autoria Própria

Tabela 5.4: Melhores resultados dos Experimentos Individuais

	P-TMDb	**P-TMDb(-)**	**P-TMDb(+)**
Classificador	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48
Experimento	EI_05	EI_05	EI_09
Características	TF-IDF	TF-IDF	Dicionário de termos frequentes
Precisão	0,477	0,474	0,577
Revocação	0,480	0,476	0,577
Medida-F	0,478	0,475	0,577

Fonte: Autoria Própria

tamanho da base original, trouxe melhorias para a qualidade da classificação como um todo. A Tabela 5.4 mostra um resumo dos melhores resultados para os Experimentos Individuais, indicando o classificador que os obteve, o experimento no qual foram obtidos, a natureza dos conjuntos de características que compuseram o experimento, e as métricas obtidas após sua realização.

5.2 EXPERIMENTOS COMBINATORIAIS

Nesta subseção apresentaremos os resultados dos Experimentos Combinatoriais, que visaram explorar a combinação entre grupos de características diferentes para o aprimoramento da classificação realizada. Duas estratégias distintas de combinação foram utilizadas: Fusão Imediata (*early fusion*), na qual as características extraídas de cada grupo foram combinadas e então utilizadas pelos algoritmos classificadores, e Fusão Tardia (*late fusion*), na qual cada classificador apresenta uma anotação para cada entrada da base e então tais classificações são combinadas usando-se os níveis de confiança que cada classificador atribui a cada um dos rótulos na classificação de cada instância.

A Tabela 5.5 apresenta os valores de precisão, revocação e medida-F para todos os Experimentos Combinatoriais realizados na base P-TMDb usando o método de Fusão Imediata, enquanto a Figura 5.4 apresenta a visualização dos valores de medida-F obtidos nesses experimentos.

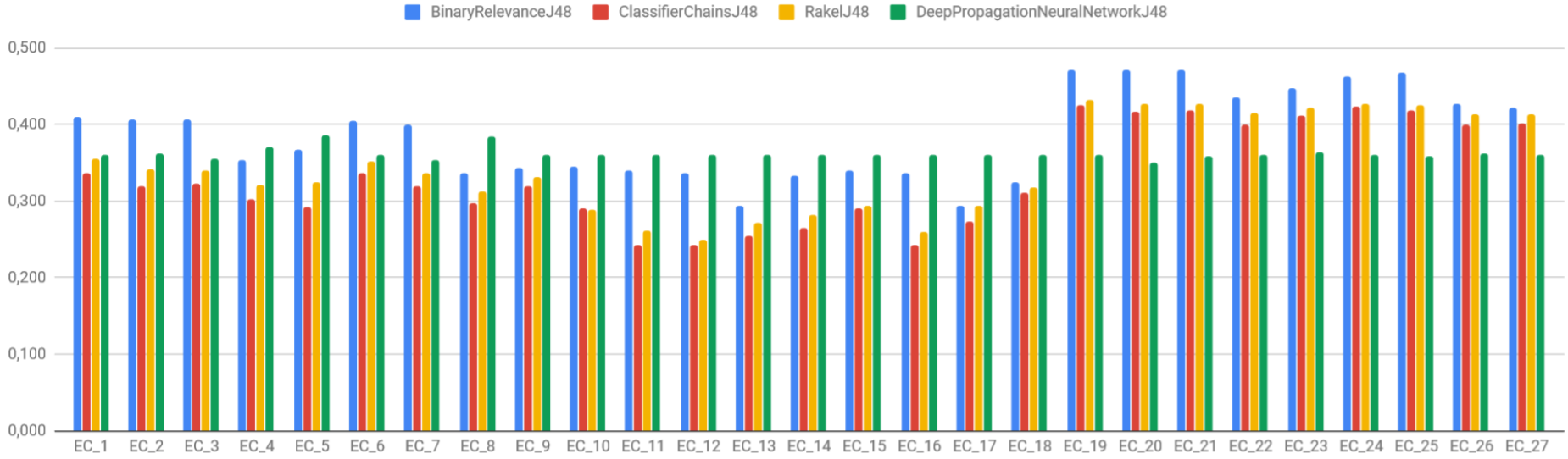
Ao comparar os resultados apresentados na Tabela 5.5 e na Figura 5.4 com aqueles apresentados na Tabela 5.1 e na Figura 5.1 referentes aos Experimentos Individuais na mesma base, é possível observar que a combinação de diferentes grupos de características teve um impacto positivo, pois obteve-se valores maiores de medida-F se comparados aos experimentos individuais.

Tabela 5.5: Resultados dos Experimentos Combinatoriais usando Fusão Imediata na base P-TMDb

Exp.	BinaryRelevanceJ48			ClassifierChainsJ48			RakelJ48			DeepPropagationNeuralNetworkJ48		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
EC_1	0,408	0,412	0,410	0,416	0,282	0,336	0,324	0,392	0,355	0,317	0,419	0,361
EC_2	0,402	0,411	0,406	0,408	0,261	0,319	0,303	0,393	0,342	0,317	0,419	0,361
EC_3	0,404	0,409	0,406	0,419	0,264	0,324	0,302	0,389	0,340	0,343	0,371	0,355
EC_4	0,351	0,355	0,353	0,328	0,279	0,302	0,271	0,395	0,322	0,345	0,406	0,371
EC_5	0,366	0,369	0,368	0,344	0,253	0,292	0,285	0,379	0,325	0,381	0,391	0,386
EC_6	0,404	0,406	0,405	0,404	0,289	0,337	0,315	0,397	0,352	0,317	0,419	0,361
EC_7	0,397	0,401	0,399	0,401	0,265	0,319	0,295	0,392	0,337	0,344	0,365	0,354
EC_8	0,335	0,338	0,336	0,312	0,283	0,297	0,265	0,380	0,312	0,374	0,396	0,385
EC_9	0,343	0,344	0,343	0,337	0,305	0,320	0,284	0,398	0,331	0,317	0,419	0,361
EC_10	0,341	0,348	0,344	0,334	0,256	0,290	0,246	0,349	0,288	0,317	0,419	0,361
EC_11	0,332	0,348	0,340	0,325	0,195	0,243	0,217	0,328	0,261	0,317	0,419	0,361
EC_12	0,325	0,348	0,336	0,324	0,194	0,242	0,207	0,315	0,250	0,317	0,419	0,361
EC_13	0,292	0,296	0,294	0,270	0,242	0,255	0,225	0,346	0,272	0,317	0,419	0,361
EC_14	0,328	0,337	0,333	0,302	0,237	0,265	0,236	0,350	0,282	0,317	0,419	0,361
EC_15	0,335	0,344	0,339	0,325	0,263	0,291	0,250	0,359	0,295	0,317	0,419	0,361
EC_16	0,332	0,340	0,336	0,311	0,199	0,243	0,214	0,328	0,259	0,317	0,419	0,361
EC_17	0,292	0,295	0,293	0,281	0,267	0,274	0,246	0,364	0,294	0,317	0,419	0,361
EC_18	0,324	0,325	0,324	0,322	0,301	0,311	0,270	0,387	0,318	0,317	0,419	0,361
EC_19	0,471	0,472	0,472	0,499	0,370	0,425	0,442	0,421	0,432	0,317	0,419	0,361
EC_20	0,471	0,472	0,472	0,487	0,364	0,417	0,430	0,425	0,428	0,351	0,350	0,351
EC_21	0,471	0,472	0,471	0,492	0,365	0,419	0,433	0,422	0,427	0,326	0,402	0,359
EC_22	0,433	0,437	0,435	0,431	0,374	0,400	0,384	0,452	0,415	0,317	0,419	0,361
EC_23	0,447	0,450	0,448	0,454	0,376	0,412	0,403	0,444	0,422	0,326	0,414	0,364
EC_24	0,462	0,464	0,463	0,487	0,374	0,423	0,426	0,428	0,427	0,317	0,419	0,361
EC_25	0,467	0,469	0,468	0,479	0,372	0,419	0,422	0,430	0,426	0,320	0,412	0,359
EC_26	0,426	0,429	0,428	0,417	0,382	0,399	0,377	0,457	0,413	0,318	0,420	0,362
EC_27	0,421	0,423	0,422	0,418	0,387	0,402	0,374	0,463	0,413	0,317	0,419	0,361

Fonte: Autoria Própria

Figura 5.4: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Imediata na base P-TMDb



Fonte: Autoria Própria

O melhor resultado, um valor de medida-F de $0,472$, foi obtido nos experimento EC_19 e EC_20 usando o classificador *BinaryRelevanceJ48*, que, como nos Experimentos Individuais, mostrou-se como o melhor classificador em quase todos os experimentos. Nota-se que os experimentos que obtiveram métricas próximas àquelas observadas no melhor caso dos Experimentos Individuais foram EC_19, EC_20, EC_21, EC_24 e EC_25, sendo que todos são referentes àqueles que implementam conjuntos de características baseados em TF-IDF. Isso mostra que a combinação conjunto de características TF-IDF com os demais não foi capaz de produzir resultados superiores ao uso do conjunto de forma isolada. Mesmo que o tamanho dos vetores de características fosse aumentado, trazendo mais informações sobre a sinopse, tais informações não puderam melhorar a qualidade da classificação final.

Outro aspecto que pode ser observado é que três dos classificadores utilizados seguiram a mesma tendência em relação a seus resultados, tendo-os acrescidos ou reduzidos sempre nos mesmos experimentos. A exceção foi o classificador *DeepPropagationNeuralNetworkJ48*, que obteve resultados muito similares em todos os experimentos, o que pode indicar que não houveram dados suficientes para que a rede neural do classificador pudesse absorver informações suficientes sobre a base.

A Tabela 5.6 mostra os resultados obtidos com a realização dos mesmos Experimentos Combinatoriais utilizando a base P-TMDb(-), sendo que a Figura 5.5 mostra uma comparação dos valores de medida-F obtidos pelos classificadores.

Paralelamente ao que pôde ser observado nos Experimentos Individuais, ao analisar a Tabela 5.6 e a Figura 5.5 nota-se que o uso do algoritmo de rebalanceamento LP-RUS não apresentou resultados significativos na classificação final. Como o que foi constatado no experimento utilizando a base de dados original, o melhor resultado experimental com a base P-TMDb(-) foi obtido no EC_19, que neste caso obteve uma medida-F de $0,470$, ligeiramente inferior ao melhor valor obtido com a base T-PMDb. Os demais resultados desses experimentos se mostraram similares aos observados na Tabela 5.5 e na Figura 5.4, o que indica que a redução de cerca de 25% da base de dados pela redução de seus *labelsets* mais frequentes não apresentou mudanças significativas na base como um todo.

Os resultados dos Experimentos Combinatoriais usando Fusão Imediata para a base rebalanceada P-TMMDb(+) são mostrados na Tabela 5.7, enquanto a Figura 5.6 apresenta um comparativo dos valores médios de medida-F obtidos nos mesmos experimentos. .

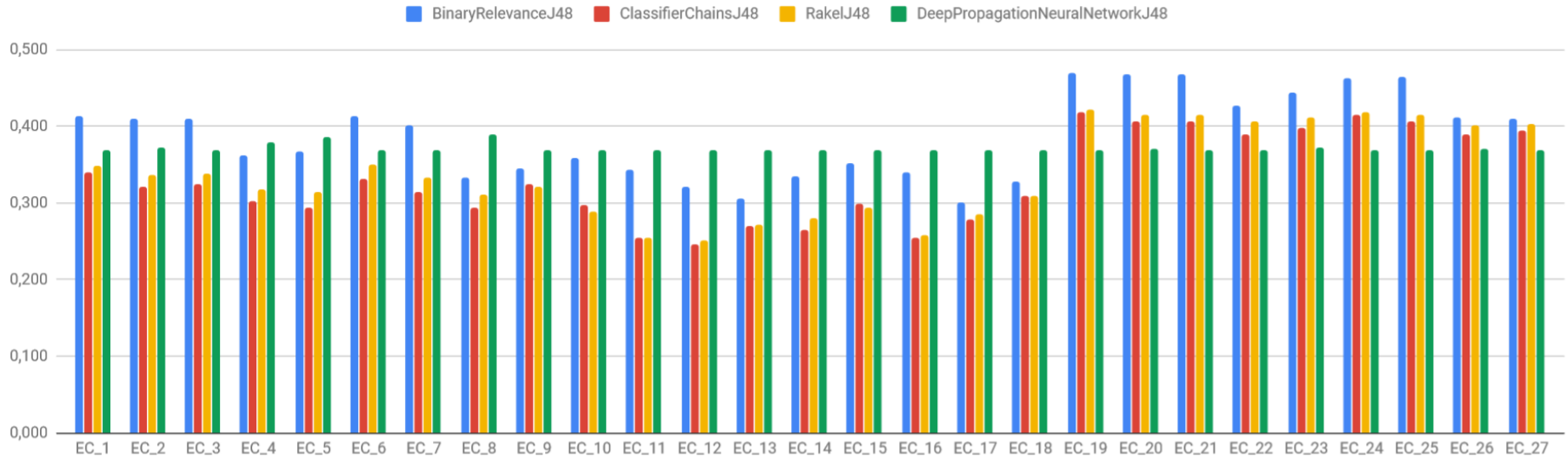
Ao se observar a Tabela 5.7 e a Figura 5.6 pode-se perceber uma melhora na classificação proporcionada pelo uso da base P-TMDb(+). Os resultados dos três primeiros classificadores foram melhorados de forma que o melhor resultado obtido do experimento

Tabela 5.6: Resultados dos Experimentos Combinatoriais usando Fusão Imediata na base P-TMDb(-)

Exp.	BinaryRelevanceJ48			ClassifierChainsJ48			RakelJ48			DeepPropagationNeuralNetworkJ48		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
EC_1	0,411	0,414	0,413	0,429	0,282	0,340	0,323	0,378	0,349	0,322	0,434	0,369
EC_2	0,408	0,413	0,410	0,424	0,258	0,321	0,301	0,382	0,337	0,338	0,418	0,372
EC_3	0,407	0,413	0,410	0,432	0,260	0,325	0,303	0,383	0,338	0,322	0,434	0,370
EC_4	0,360	0,365	0,363	0,333	0,278	0,303	0,271	0,385	0,318	0,360	0,405	0,379
EC_5	0,365	0,370	0,367	0,354	0,251	0,293	0,282	0,356	0,315	0,377	0,395	0,386
EC_6	0,413	0,415	0,414	0,413	0,277	0,332	0,320	0,389	0,351	0,322	0,434	0,369
EC_7	0,399	0,405	0,402	0,408	0,256	0,314	0,295	0,383	0,333	0,322	0,434	0,370
EC_8	0,330	0,335	0,332	0,313	0,276	0,293	0,268	0,371	0,311	0,374	0,407	0,389
EC_9	0,345	0,346	0,346	0,347	0,305	0,325	0,280	0,377	0,321	0,322	0,434	0,369
EC_10	0,347	0,371	0,358	0,360	0,253	0,297	0,250	0,340	0,288	0,322	0,434	0,369
EC_11	0,335	0,351	0,343	0,346	0,202	0,255	0,215	0,311	0,254	0,322	0,434	0,369
EC_12	0,312	0,329	0,320	0,333	0,195	0,245	0,211	0,311	0,251	0,322	0,434	0,369
EC_13	0,303	0,307	0,305	0,289	0,253	0,270	0,226	0,341	0,272	0,322	0,434	0,369
EC_14	0,330	0,339	0,334	0,314	0,229	0,265	0,241	0,337	0,281	0,322	0,434	0,369
EC_15	0,346	0,358	0,351	0,349	0,261	0,299	0,254	0,349	0,294	0,322	0,434	0,369
EC_16	0,331	0,350	0,340	0,338	0,204	0,254	0,216	0,318	0,257	0,322	0,434	0,369
EC_17	0,299	0,303	0,301	0,293	0,267	0,279	0,241	0,350	0,285	0,322	0,434	0,369
EC_18	0,327	0,329	0,328	0,327	0,293	0,309	0,267	0,370	0,310	0,322	0,434	0,369
EC_19	0,469	0,470	0,470	0,492	0,363	0,418	0,440	0,404	0,421	0,322	0,434	0,369
EC_20	0,467	0,470	0,468	0,477	0,354	0,406	0,426	0,405	0,415	0,371	0,371	0,371
EC_21	0,463	0,474	0,468	0,482	0,352	0,407	0,428	0,403	0,415	0,322	0,434	0,370
EC_22	0,424	0,428	0,426	0,418	0,365	0,390	0,378	0,440	0,407	0,322	0,434	0,369
EC_23	0,444	0,445	0,444	0,445	0,361	0,398	0,399	0,426	0,412	0,336	0,420	0,372
EC_24	0,462	0,464	0,463	0,477	0,366	0,414	0,424	0,414	0,419	0,322	0,434	0,369
EC_25	0,462	0,467	0,464	0,469	0,360	0,407	0,416	0,414	0,415	0,322	0,434	0,369
EC_26	0,411	0,413	0,412	0,405	0,375	0,390	0,368	0,442	0,402	0,324	0,433	0,370
EC_27	0,409	0,410	0,409	0,413	0,379	0,395	0,369	0,445	0,403	0,322	0,434	0,369

Fonte: Autoria Própria

Figura 5.5: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Imediata na base P-TMDb(-)



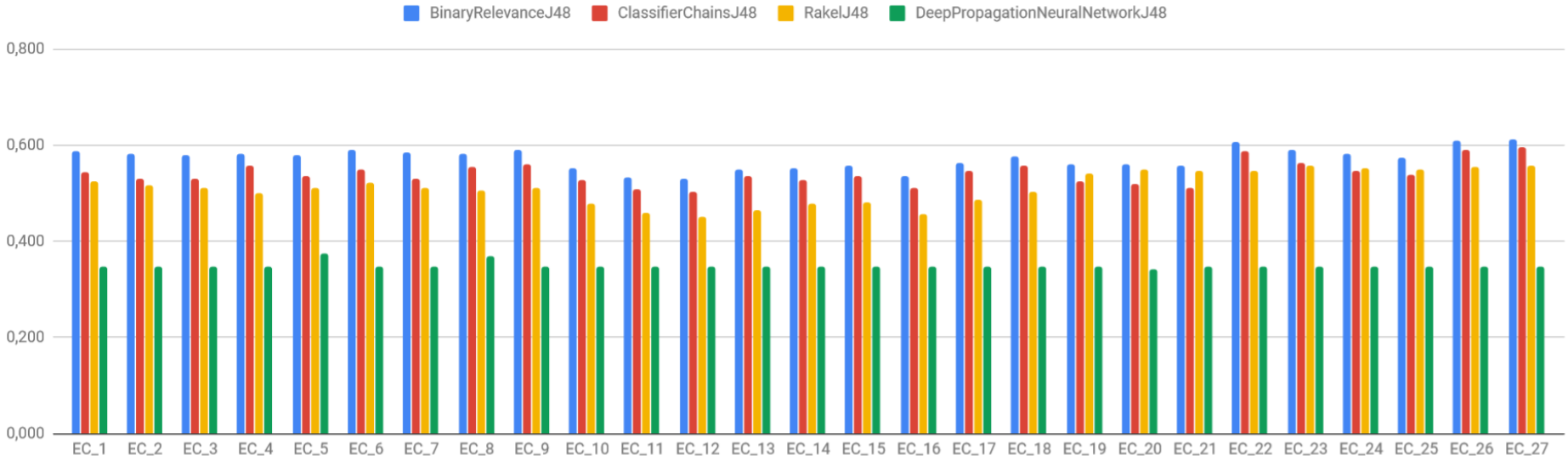
Fonte: Autoria Própria

Tabela 5.7: Resultados dos Experimentos Combinatoriais usando Fusão Imediata na base P-TMDb(+)

Exp.	BinaryRelevanceJ48			ClassifierChainsJ48			RakelJ48			DeepPropagationNeuralNetworkJ48		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
EC_1	0,587	0,587	0,587	0,566	0,521	0,543	0,482	0,575	0,524	0,324	0,372	0,347
EC_2	0,582	0,584	0,583	0,559	0,505	0,530	0,466	0,578	0,516	0,324	0,372	0,347
EC_3	0,580	0,582	0,581	0,560	0,504	0,531	0,464	0,571	0,512	0,326	0,370	0,346
EC_4	0,579	0,583	0,581	0,549	0,565	0,557	0,444	0,574	0,501	0,324	0,372	0,347
EC_5	0,579	0,581	0,580	0,556	0,516	0,535	0,472	0,556	0,510	0,368	0,380	0,374
EC_6	0,590	0,592	0,591	0,564	0,536	0,550	0,476	0,577	0,522	0,324	0,372	0,347
EC_7	0,583	0,585	0,584	0,552	0,513	0,531	0,461	0,577	0,512	0,326	0,369	0,346
EC_8	0,580	0,586	0,583	0,552	0,555	0,553	0,456	0,565	0,504	0,358	0,383	0,369
EC_9	0,587	0,592	0,589	0,557	0,565	0,561	0,466	0,566	0,511	0,324	0,372	0,347
EC_10	0,549	0,554	0,551	0,522	0,534	0,528	0,421	0,549	0,477	0,324	0,372	0,347
EC_11	0,531	0,536	0,533	0,506	0,513	0,510	0,399	0,539	0,459	0,324	0,372	0,347
EC_12	0,528	0,532	0,530	0,502	0,503	0,503	0,391	0,529	0,450	0,324	0,372	0,347
EC_13	0,547	0,550	0,549	0,521	0,550	0,535	0,403	0,549	0,464	0,324	0,372	0,347
EC_14	0,548	0,553	0,551	0,528	0,528	0,528	0,428	0,541	0,478	0,324	0,372	0,347
EC_15	0,557	0,559	0,558	0,528	0,544	0,536	0,424	0,554	0,480	0,324	0,372	0,347
EC_16	0,534	0,537	0,535	0,505	0,514	0,510	0,396	0,538	0,456	0,324	0,372	0,347
EC_17	0,559	0,565	0,562	0,537	0,555	0,546	0,433	0,556	0,487	0,324	0,372	0,347
EC_18	0,574	0,580	0,577	0,550	0,567	0,558	0,451	0,566	0,502	0,324	0,372	0,347
EC_19	0,561	0,561	0,561	0,570	0,485	0,524	0,544	0,539	0,542	0,324	0,372	0,347
EC_20	0,557	0,560	0,559	0,567	0,481	0,521	0,534	0,568	0,550	0,342	0,342	0,342
EC_21	0,550	0,565	0,557	0,564	0,468	0,511	0,536	0,556	0,546	0,324	0,372	0,346
EC_22	0,604	0,607	0,606	0,580	0,594	0,587	0,501	0,599	0,546	0,324	0,372	0,347
EC_23	0,589	0,592	0,591	0,589	0,537	0,562	0,541	0,578	0,559	0,326	0,370	0,347
EC_24	0,582	0,584	0,583	0,576	0,522	0,547	0,528	0,576	0,551	0,324	0,372	0,347
EC_25	0,572	0,577	0,575	0,573	0,510	0,539	0,522	0,579	0,549	0,324	0,372	0,347
EC_26	0,608	0,613	0,610	0,586	0,596	0,591	0,519	0,596	0,555	0,325	0,372	0,347
EC_27	0,609	0,614	0,611	0,589	0,601	0,595	0,519	0,600	0,557	0,324	0,372	0,347

Fonte: Autoria Própria

Figura 5.6: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Imediata na base P-TMDb(+)



Fonte: Autoria Própria

Tabela 5.8: Melhores resultados dos Experimentos Combinatoriais usando Fusão Imediata

	P-TMDb	**P-TMDb(-)**	**P-TMDb(+)**
Classificador	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48
Experimento	EC_19 e EC_20	EC_19	EC_27
Características	Estruturais, TF-IDF, Nome do gênero	Estruturais, TF-IDF, Nome do gênero	Estruturais, TF-IDF, Nome do gênero, Classes Gramaticais, Aspectos Linguísticos, LIWC, LDA
Precisão	0,471	0,469	0,609
Revocação	0,472	0,470	0,614
Medida-F	0,472	0,470	0,611

Fonte: Autoria Própria

EC_27 chegou a uma medida-F de $0,611$ usando o classificador *BinaryRelevanceJ48*, sendo que os dois outros melhores experimentos, EC_26 e EC_22 apresentaram resultados de $0,610$ e $0,606$ de medida-F, respectivamente.

Nota-se que os três experimentos que apresentaram os melhores resultados usaram combinações dos grupos de características 2 e 7, que correspondem aos conjuntos de características baseadas em TF-IDF e à contagem normalizada de classes do LIWC, respectivamente. Tal observação é condizente com o que já foi constatado anteriormente e reforça a conclusão de que essas estratégias de extração de características são adequadas para o domínio estudado.

Percebe-se também que, enquanto o classificador *BinaryRelevanceJ48* mostra-se mais uma vez como o mais adequado para a classificação realizada, o classificador *ClassifierChainsJ48* passou a ter, na maioria dos casos, o segundo maior resultado de classificação, enquanto que nos testes com a base original ele se mostrou frequentemente como o pior entre os classificadores estudados. Pode-se concluir então que o rebalanceamento da base original foi capaz de alterar a base P-TMDb para que essa se tornasse mais adequada para o uso da estratégia de cadeias de classificadores. Por fim, nota-se que mais uma vez não houve alteração significativa dos resultados do classificador *DeepPropagationNeuralNetworkJ48*, mostrando que o algoritmo LP-RUS não foi capaz de produzir informações suficientes para que a rede neural pudesse ser efetivamente treinada para a base estudada.

A Tabela 5.8 apresenta os melhores resultados obtidos para todas as execuções dos Experimentos Combinatoriais usando Fusão Imediata, listando os classificadores em que a melhor classificação foi obtida, o experimento que obteve o melhor resultado, a natureza dos conjuntos de características utilizados em tal experimento, e as métricas que compõem os melhores resultados entre as bases.

Nos experimentos com Fusão Tardia, três estratégias de fusão baseadas nos níveis de confiança da saída de cada classificador foram utilizadas: *avg*, na qual a média entre todos os níveis de confiança foi considerada, *max*, na qual a classificação de cada entrada foi obtida usando-se do valor máximo de confiança entre os classificadores, e *sum*, na qual todas as confianças dos classificadores individuais foram somadas para gerar uma nova confiança para a classificação. Estratégias utilizando o valor mínimo e o produto entre todas as confianças foram exploradas, mas devido à várias instâncias em que a confiança produzida por um determinado classificador foi atribuída como 0, tais estratégias não produziram resultados.

Devido a restrições de tempo e ao fato de seus resultados não terem mostrado diferenças significativas em relação aos resultados da base original, a base P-TMDb(-) não foi utilizada nos Experimentos Combinatoriais usando Fusão Tardia.

As Tabelas 5.9, 5.10 e 5.11 apresentam os resultados dos Experimentos Combinatoriais usando das três estratégias de Fusão Tardia com a base P-TMDb. As Figuras 5.7, 5.8 e 5.7 apresentam visualizações para os valores de medida-F obtidos para as estratégias de fusão média, máxima e soma, respectivamente.

Ao observar as Tabelas 5.9, 5.10 e 5.11 e as Figuras 5.7, 5.8 e 5.7 nota-se que, em relação aos melhores resultados obtidos pelo classificador *BinaryRelevanceJ48*, uso das estratégias *max* e *avg* mostraram tendências similares, com valores de medida-F de $0,481$ e $0,478$, respectivamente, para suas melhores classificações. A estratégia *sum*, no entanto, teve um comportamento diferente, com o classificador *ClassifierChainsJ48* obtendo um valor medida-F de $0,454$ como melhor resultado.

Ao comparar os resultados da Figura 5.8 com os da Figura 5.4, é possível constatar que a estratégia de Fusão Tardia melhorou a maioria dos resultados, aprimorando mesmo alguns casos do classificador *DeepPropagationNeuralNetworkJ48* que previamente não havia sofrido alterações significativas. Fato similar pode ser constatado ao se comparar as Figuras 5.7 e 5.4, que apresentam as mesmas tendências de melhoramento, mas com valores ligeiramente menores. No entanto, quando se compara os resultados das Figuras 5.9 e 5.4, nota-se casos em que classificações específicas foram melhoradas e pioradas, indicando que tal estratégia não seja apropriada para utilização no caso geral do contexto estudado.

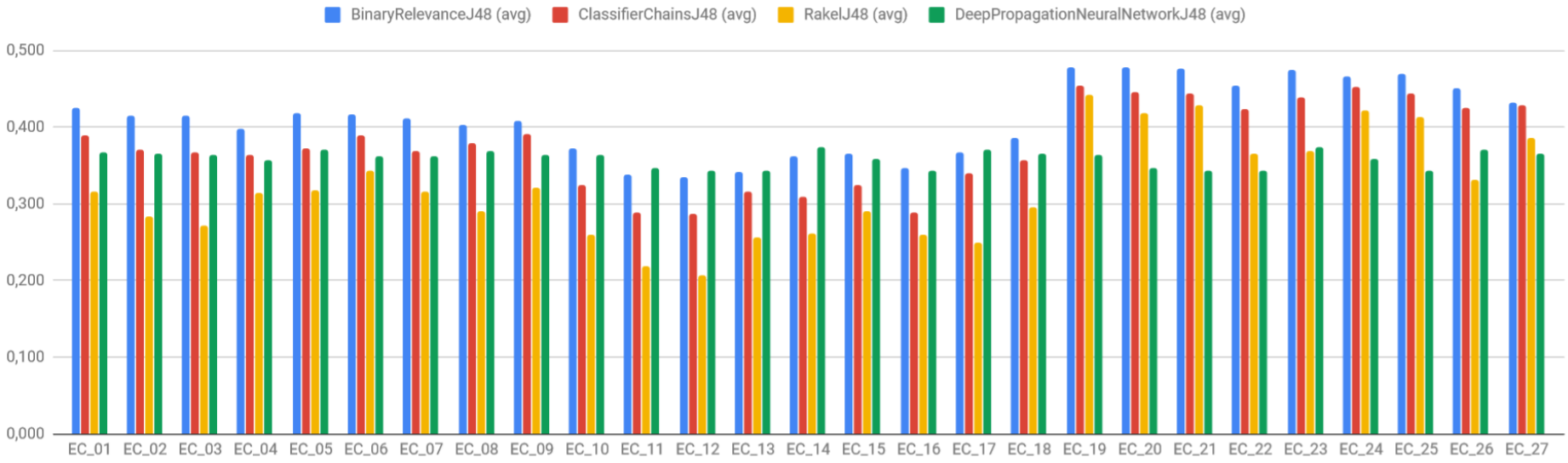
Por fim, as Tabelas 5.12, 5.13 e 5.14 mostram todos os resultados da execução dos Experimentos Combinatoriais usando estratégias de Fusão Tardia com base rebalanceada P-TMDb(+), enquanto que as Figuras 5.10, 5.11 e 5.12 apresentam a comparação dos valores de medida-F para as estratégias de Fusão Tardia *avg*, *max* e *sum*, respectivamente.

Tabela 5.9: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das médias de confiança na base P-TMDb

Exp.	BinaryRelevanceJ48	ClassifierChainsJ48	RakelJ48	DeepPropagationNeuralNetworkJ48
EC_01	0,425	0,390	0,315	0,367
EC_02	0,415	0,370	0,283	0,365
EC_03	0,415	0,368	0,272	0,363
EC_04	0,398	0,365	0,315	0,357
EC_05	0,418	0,372	0,318	0,371
EC_06	0,416	0,390	0,343	0,362
EC_07	0,411	0,369	0,316	0,363
EC_08	0,403	0,380	0,291	0,370
EC_09	0,409	0,391	0,320	0,364
EC_10	0,372	0,324	0,260	0,364
EC_11	0,339	0,289	0,219	0,347
EC_12	0,335	0,287	0,208	0,343
EC_13	0,342	0,315	0,257	0,343
EC_14	0,361	0,308	0,261	0,373
EC_15	0,366	0,325	0,291	0,359
EC_16	0,346	0,289	0,259	0,343
EC_17	0,367	0,339	0,249	0,370
EC_18	0,386	0,358	0,295	0,365
EC_19	0,478	0,454	0,442	0,364
EC_20	0,478	0,446	0,419	0,347
EC_21	0,476	0,444	0,429	0,343
EC_22	0,454	0,423	0,365	0,343
EC_23	0,474	0,439	0,368	0,373
EC_24	0,467	0,453	0,422	0,359
EC_25	0,469	0,443	0,414	0,343
EC_26	0,451	0,426	0,331	0,370
EC_27	0,432	0,428	0,386	0,365

Fonte: Autoria Própria

Figura 5.7: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das médias de confiança na base P-TMDb



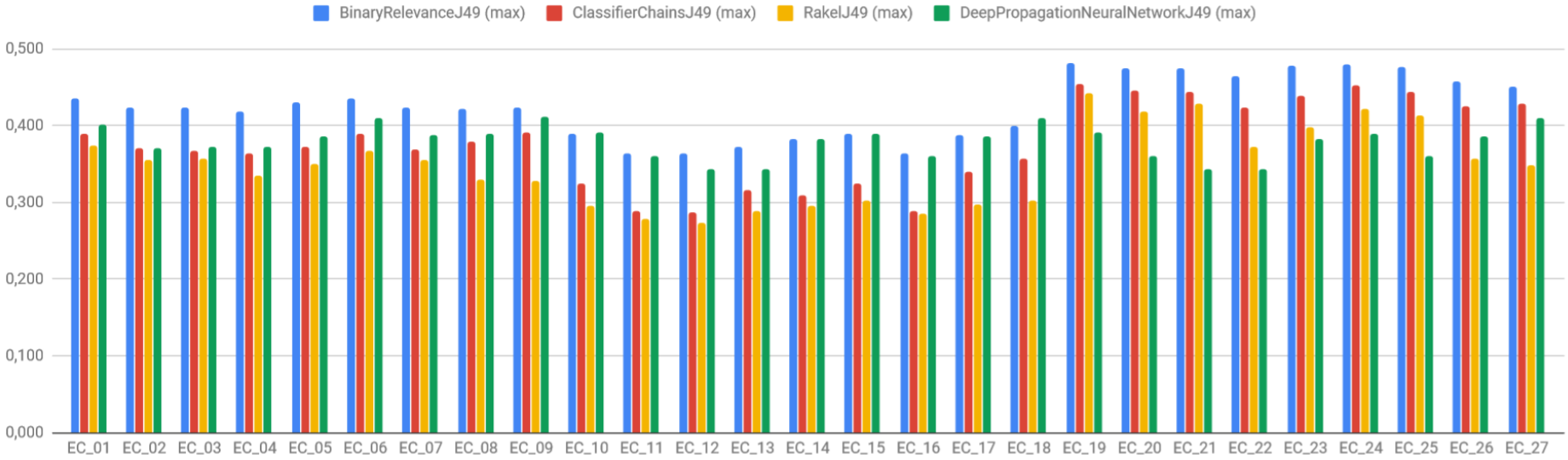
Fonte: Autoria Própria

Tabela 5.10: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das confianças máximas na base P-TMDb

Exp.	BinaryRelevanceJ48	ClassifierChainsJ48	RakelJ48	DeepPropagationNeuralNetworkJ48
EC_01	0,435	0,390	0,373	0,401
EC_02	0,423	0,370	0,355	0,371
EC_03	0,423	0,368	0,357	0,373
EC_04	0,418	0,365	0,335	0,373
EC_05	0,430	0,372	0,350	0,387
EC_06	0,436	0,390	0,367	0,410
EC_07	0,423	0,369	0,355	0,388
EC_08	0,422	0,380	0,329	0,389
EC_09	0,423	0,391	0,328	0,411
EC_10	0,389	0,324	0,295	0,391
EC_11	0,364	0,289	0,279	0,361
EC_12	0,363	0,287	0,273	0,343
EC_13	0,372	0,315	0,288	0,343
EC_14	0,383	0,308	0,296	0,383
EC_15	0,390	0,325	0,302	0,390
EC_16	0,364	0,289	0,285	0,361
EC_17	0,388	0,339	0,298	0,385
EC_18	0,400	0,358	0,302	0,410
EC_19	0,481	0,454	0,442	0,391
EC_20	0,475	0,446	0,419	0,361
EC_21	0,475	0,444	0,429	0,343
EC_22	0,464	0,423	0,373	0,343
EC_23	0,478	0,439	0,397	0,383
EC_24	0,480	0,453	0,422	0,390
EC_25	0,476	0,443	0,414	0,361
EC_26	0,457	0,426	0,356	0,385
EC_27	0,451	0,428	0,348	0,410

Fonte: Autoria Própria

Figura 5.8: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das confianças máximas na base P-TMDb



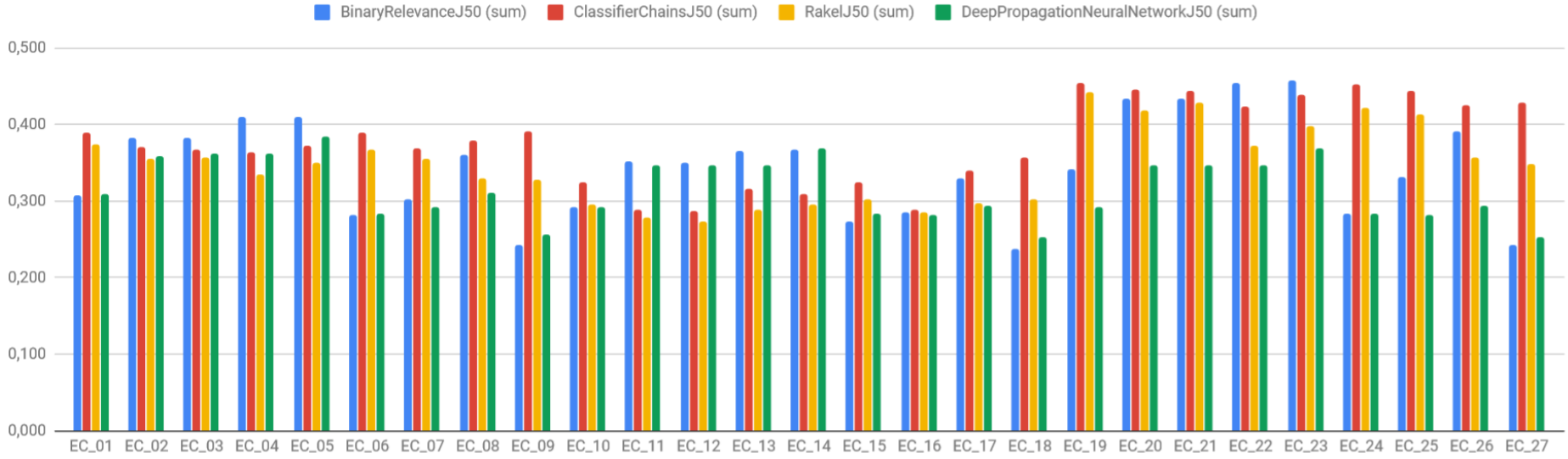
Fonte: Autoria Própria

Tabela 5.11: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das soma das confianças na base P-TMDb

Exp.	BinaryRelevanceJ48	ClassifierChainsJ48	RakelJ48	DeepPropagationNeuralNetworkJ48
EC_01	0,307	0,390	0,373	0,309
EC_02	0,383	0,370	0,355	0,359
EC_03	0,383	0,368	0,357	0,362
EC_04	0,410	0,365	0,335	0,362
EC_05	0,410	0,372	0,350	0,384
EC_06	0,282	0,390	0,367	0,284
EC_07	0,302	0,369	0,355	0,292
EC_08	0,361	0,380	0,329	0,310
EC_09	0,242	0,391	0,328	0,256
EC_10	0,292	0,324	0,295	0,292
EC_11	0,351	0,289	0,279	0,347
EC_12	0,351	0,287	0,273	0,347
EC_13	0,365	0,315	0,288	0,347
EC_14	0,367	0,308	0,296	0,369
EC_15	0,273	0,325	0,302	0,283
EC_16	0,286	0,289	0,285	0,282
EC_17	0,330	0,339	0,298	0,295
EC_18	0,237	0,358	0,302	0,253
EC_19	0,342	0,454	0,442	0,292
EC_20	0,433	0,446	0,419	0,347
EC_21	0,433	0,444	0,429	0,347
EC_22	0,454	0,423	0,373	0,347
EC_23	0,458	0,439	0,397	0,369
EC_24	0,283	0,453	0,422	0,283
EC_25	0,331	0,443	0,414	0,282
EC_26	0,392	0,426	0,356	0,295
EC_27	0,243	0,428	0,348	0,253

Fonte: Autoria Própria

Figura 5.9: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das soma das confianças na base P-TMDb



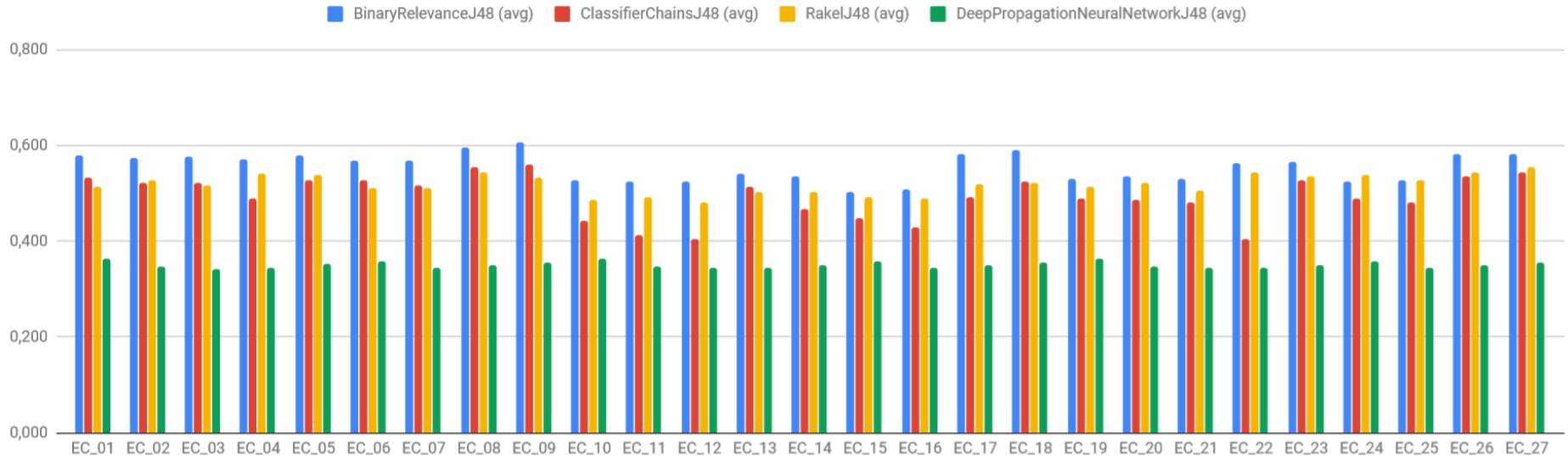
Fonte: Autoria Própria

Tabela 5.12: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das médias de confiança na base P-TMDb(+)

Exp.	BinaryRelevanceJ48	ClassifierChainsJ48	RakelJ48	DeepPropagationNeuralNetworkJ48
EC_01	0,578	0,532	0,515	0,363
EC_02	0,575	0,522	0,528	0,347
EC_03	0,575	0,522	0,517	0,342
EC_04	0,570	0,490	0,541	0,344
EC_05	0,579	0,529	0,537	0,352
EC_06	0,569	0,526	0,510	0,357
EC_07	0,567	0,517	0,511	0,344
EC_08	0,596	0,555	0,543	0,350
EC_09	0,607	0,560	0,533	0,355
EC_10	0,526	0,443	0,486	0,363
EC_11	0,526	0,413	0,491	0,347
EC_12	0,526	0,404	0,481	0,344
EC_13	0,540	0,515	0,504	0,344
EC_14	0,535	0,468	0,503	0,351
EC_15	0,502	0,449	0,493	0,357
EC_16	0,509	0,430	0,489	0,344
EC_17	0,582	0,492	0,519	0,351
EC_18	0,590	0,524	0,522	0,355
EC_19	0,531	0,490	0,514	0,363
EC_20	0,535	0,486	0,522	0,347
EC_21	0,529	0,481	0,506	0,344
EC_22	0,564	0,404	0,545	0,344
EC_23	0,567	0,527	0,536	0,351
EC_24	0,524	0,490	0,537	0,357
EC_25	0,527	0,482	0,527	0,344
EC_26	0,583	0,535	0,543	0,351
EC_27	0,583	0,543	0,556	0,355

Fonte: Autoria Própria

Figura 5.10: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das médias de confiança na base P-TMDb(+)



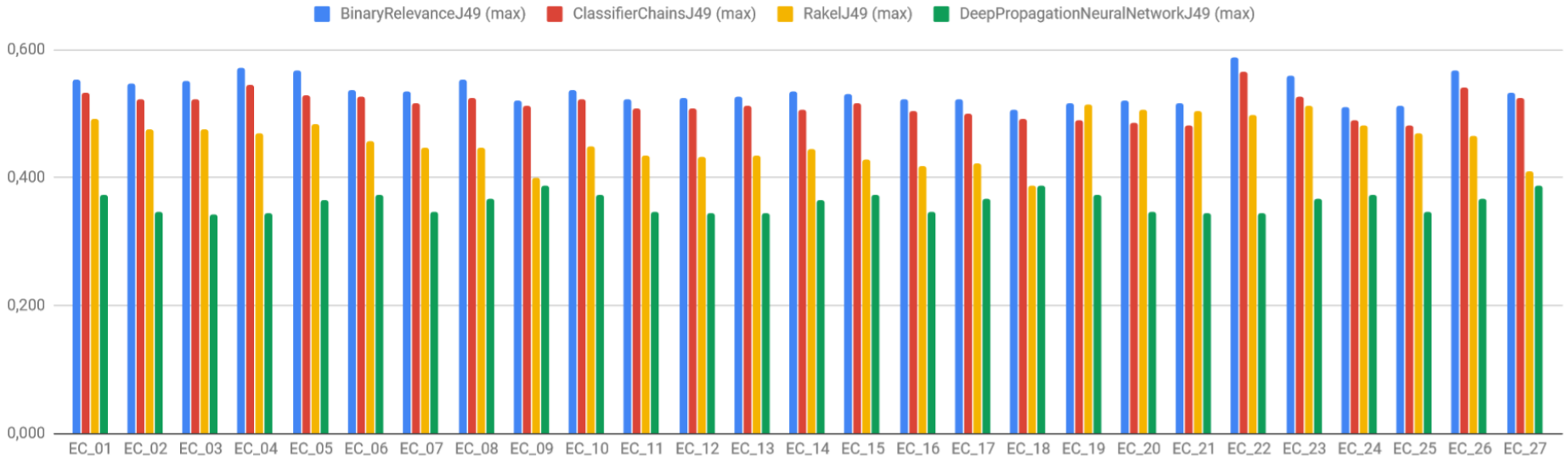
Fonte: Autoria Própria

Tabela 5.13: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das confianças máximas na base P-TMDb(+)

Exp.	BinaryRelevanceJ48	ClassifierChainsJ48	RakelJ48	DeepPropagationNeuralNetworkJ48
EC_01	0,553	0,532	0,492	0,372
EC_02	0,547	0,522	0,476	0,347
EC_03	0,551	0,522	0,475	0,342
EC_04	0,571	0,544	0,468	0,344
EC_05	0,567	0,529	0,483	0,366
EC_06	0,537	0,526	0,458	0,374
EC_07	0,534	0,517	0,447	0,347
EC_08	0,554	0,525	0,446	0,366
EC_09	0,521	0,511	0,400	0,388
EC_10	0,538	0,522	0,449	0,372
EC_11	0,523	0,509	0,435	0,347
EC_12	0,525	0,509	0,433	0,344
EC_13	0,527	0,513	0,434	0,344
EC_14	0,535	0,506	0,444	0,366
EC_15	0,530	0,516	0,429	0,373
EC_16	0,521	0,505	0,417	0,347
EC_17	0,522	0,500	0,423	0,366
EC_18	0,506	0,491	0,388	0,388
EC_19	0,516	0,490	0,514	0,372
EC_20	0,521	0,486	0,506	0,347
EC_21	0,517	0,481	0,503	0,344
EC_22	0,588	0,566	0,497	0,345
EC_23	0,559	0,527	0,511	0,366
EC_24	0,510	0,490	0,481	0,373
EC_25	0,512	0,482	0,469	0,347
EC_26	0,566	0,540	0,465	0,367
EC_27	0,532	0,524	0,409	0,388

Fonte: Autoria Própria

Figura 5.11: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das confianças máximas na base P-TMDb(+)



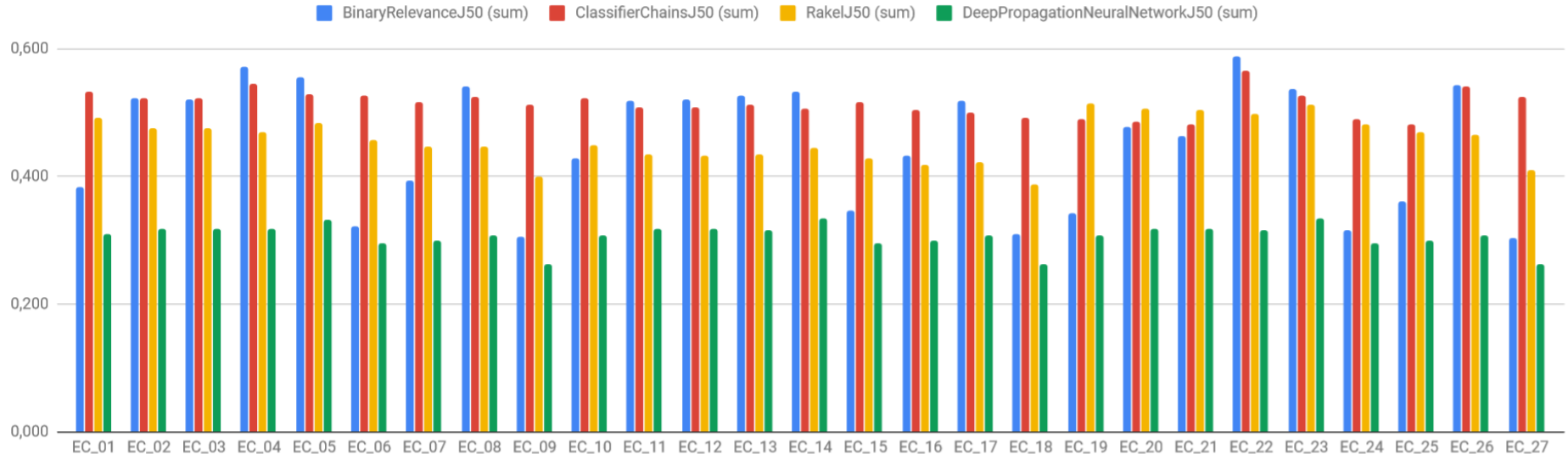
Fonte: Autoria Própria

Tabela 5.14: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das soma das confianças na base P-TMDb(+)

Exp.	BinaryRelevanceJ48	ClassifierChainsJ48	RakelJ48	DeepPropagationNeuralNetworkJ48
EC_01	0,384	0,532	0,492	0,309
EC_02	0,523	0,522	0,476	0,317
EC_03	0,520	0,522	0,475	0,318
EC_04	0,571	0,544	0,468	0,317
EC_05	0,554	0,529	0,483	0,333
EC_06	0,322	0,526	0,458	0,296
EC_07	0,394	0,517	0,447	0,300
EC_08	0,542	0,525	0,446	0,308
EC_09	0,306	0,511	0,400	0,263
EC_10	0,427	0,522	0,449	0,308
EC_11	0,518	0,509	0,435	0,317
EC_12	0,520	0,509	0,433	0,317
EC_13	0,527	0,513	0,434	0,317
EC_14	0,533	0,506	0,444	0,333
EC_15	0,346	0,516	0,429	0,296
EC_16	0,431	0,505	0,417	0,300
EC_17	0,519	0,500	0,423	0,308
EC_18	0,309	0,491	0,388	0,263
EC_19	0,342	0,490	0,514	0,308
EC_20	0,477	0,486	0,506	0,317
EC_21	0,463	0,481	0,503	0,317
EC_22	0,587	0,566	0,497	0,317
EC_23	0,538	0,527	0,511	0,334
EC_24	0,316	0,490	0,481	0,296
EC_25	0,360	0,482	0,469	0,300
EC_26	0,542	0,540	0,465	0,308
EC_27	0,304	0,524	0,409	0,263

Fonte: Autoria Própria

Figura 5.12: Medida-F de cada classificador para cada Experimento Combinatorial usando Fusão Tardia das soma das confianças na base P-TMDb(+)



Fonte: Autoria Própria

Após a análise das Tabelas 5.12, 5.13 e 5.14, bem como das Figuras 5.10, 5.11 e 5.12, pode-se observar uma melhoria geral da qualidade das classificações similar ao observado quando do uso da base P-TMDb(+) nos Experimentos Combinatoriais usando Fusão Imediata. Os maiores valores de medida-F obtidos para as estratégias de fusão *avg*, *max* e *sum* foram respectivamente $0,607$, $0,588$ e $0,587$, mostrando-se ligeiramente inferiores ao melhor resultado da Fusão Imediata na mesma base de dados.

A diferença mais notável quando se compara os resultados obtidos com Fusão Imediata com os resultados das estratégias de Fusão Tardia na base P-TMDb(+) foi o fato da segunda estratégia apresentar seu melhor resultado no experimento E_09, em comparação ao resultado do experimento E_19 da base original, que, como observado anteriormente, faz com que o uso de conjuntos de características baseados em dicionários de termos mais frequentes por gênero se sobressaia em relação ao uso da medida TF-IDF.

Ao final da realização dos experimentos, foi construída uma tabela contendo os melhores resultados para todos os métodos de classificação utilizados neste estudo. A Tabela 5.15 compila os melhores resultados dos Experimentos Individuais, Experimentos Combinatoriais com Fusão Imediata e Experimentos Combinatoriais com Fusão Tardia, citando o classificador utilizado na obtenção do melhor resultado, o experimento no qual o melhor resultado foi obtido, os grupos de características utilizados em tais experimentos, e todas as métricas obtidas de suas execuções.

Ao analisar a Tabela 5.15 foi possível observar que o uso de conjuntos de características TF-IDF foi, no caso geral, a melhor estratégia para a classificação realizada neste estudo, sendo que todos os casos em que valores TF-IDF não estavam no melhor experimento, conjuntos de características que fazem uso de dicionários de termos mais frequentes por gênero foram usados em seu lugar. Também é possível notar que o rebalanceamento da base pelo algoritmo pode trazer uma melhora significativa na qualidade da classificação quando usado para expandir a base de dados, sendo que em todos os casos em que a base foi reduzida pelo algoritmo LP-RUS não foi observada alteração significativa dos resultados. Por fim, observa-se que estratégias de combinação de conjuntos de característica podem ainda ser utilizadas para proporcionar melhoras na classificação, sendo que estratégias de Fusão Imediata se mostraram mais efetivas na base rebalanceada, enquanto que as estratégias de Fusão Tardia foram mais efetivas na base original.

Tabela 5.15: Melhores resultados de todos os experimentos

	Experimentos Individuais			**Experimentos Combinatoriais (Fusão Imediata)**			**Experimentos Combinatoriais (Fusão Tardia)**	
Base	**P-TMDb**	**P-TMDb(-)**	**P-TMDb(+)**	**P-TMDb**	**P-TMDb(-)**	**P-TMDb(+)**	**P-TMDb**	**P-TMDb(+)**
Classificador	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48	BinaryRelevanceJ48
Experimento	EI_05	EI_05	EI_09	EC_19 e EC_20	EC_19	EC_27	EC_19	EC_09
Características	TF-IDF	TF-IDF	Dicionário de termos frequentes	Estruturais, TF-IDF, Nome do gênero	Estruturais, TF-IDF, Nome do gênero	Estruturais, TF-IDF, Nome do gênero, Classes Gramaticais, Aspectos Linguísticos, LIWC, LDA	Estruturais, TF-IDF, Nome do gênero	Estruturais, Nome do gênero, Dicionários, Classes Gramaticais, Aspectos Linguísticos, LIWC, LDA
Precisão	0,477	0,474	0,577	0,471	0,469	0,609	0,399	0,610
Revocação	0,480	0,476	0,577	0,472	0,470	0,614	0,605	0,609
Medida-F	0,478	0,475	0,577	0,472	0,470	0,611	0,481	0,607

Fonte: Autoria Própria

6

CONCLUSÃO

Neste documento foi apresentado um trabalho sobre a utilização de características textuais extraídas de sinopses escritas em língua portuguesa para o problema de classificação multirrótulo de gêneros cinematográficos. Diversas abordagens para tal classificação foram testadas, com experimentos fazendo uso de 9 grupos contendo conjuntos de características extraídos de diversos modos, 31 Experimentos Individuais testando o uso de cada conjunto de característica implementado neste estudo e 27 Experimentos Combinatoriais, nos quais os grupos de características foram combinados de maneira semi-exaustiva e então usados para classificar 3 bases de dados contendo sinopses na língua portuguesa: a base P-TMDb, contendo 13.394 sinopses; a base P-TMDb(-), derivada pelo rebalanceamento da base original pelo algoritmo LP-RUS e contendo 10.150 sinopses; e a base P-TMDb(+), derivada pelo rebalanceamento da base original pelo algoritmo LP-ROS e contendo 16.803 sinopses. Todos os experimentos foram conduzidos utilizando-se quatro classificadores multirrótulo, a saber: Relevância Binária, Cadeias de Classificadores, RAKEL (*Random k-labelsets*) e DBPNN (*Deep Back-Propagation Neural Networks*), sendo que todos os classificadores multirrótulo fizeram uso do classificador base J48.

Os melhores resultados obtidos com a base de dados original foram medida-F de $0,478$ para os Experimentos Individuais e $0,481$ para os Experimentos Combinatoriais, enquanto que os melhores resultados obtidos para as bases de dados derivadas por rebalanceamento foram medida-F de $0,577$ para os Experimentos Individuais e $0,611$ para os Experimentos Combinatoriais. Dentre todos os conjuntos de característica considerados, os que fizeram uso de características TF-IDF se sobressaíram na maioria dos experimentos. A mesma observação pode ser feita para os classificadores baseados em Relevância Binária, que obtiveram os

melhores resultados na maioria dos testes realizados. Tais resultados mostram-se condizentes com os apresentados na literatura, embora uma comparação direta não seja possível devido às diferenças existentes entre as bases de sinopses utilizadas.

Pôde-se também concluir que o rebalanceamento da base original trouxe melhorias significativas nos resultados quando feito pelo algoritmo LP-ROS, que criou a base P-TMDb(+) ao duplicar aleatoriamente entradas dos *labelsets* mais raros da base original até que seu tamanho fosse acrescido em torno de 25%. O mesmo não pôde ser observado quando se usou o algoritmo LP-RUS, que criou a base P-TMDb(-) ao remover aleatoriamente entradas dos *labelsets* mais comuns da base original até que se conseguisse uma redução de cerca de 25%.

A partir dos resultados deste estudo, sugere-se duas linhas principais de trabalhos futuros. A primeira delas diz respeito ao uso de informações extraídas a partir das sinopses em trabalhos voltados para bases de dados multimodais, de forma que essas informações possam ser utilizadas em conjunto com informações audiovisuais, que se mostram comuns para a área de classificação de gêneros cinematográficos. A segunda linha é a investigação de outros algoritmos e/ou estratégias de classificação multirrótulo, seja por meio de novos classificadores base para métodos de classificação multirrótulo conhecidos ou por meio de novos métodos multirrótulo. Por fim, destaca-se como um trabalho futuro a criação de uma base de dados multimodal a partir da base P-TMDB, o que viabilizaria estudos como os citados na primeira linha de trabalhos futuros com foco em língua portuguesa.

REFERÊNCIAS

AUSTIN, A. et al. Characterization of movie genre based on music score. In: IEEE. **2010 IEEE International Conference on Acoustics, Speech and Signal Processing**. [S.l.], 2010. p. 421–424.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: "O'Reilly Media, Inc.", 2009.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BRIGGS, F.; FERN, X. Z.; RAICH, R. Context-aware miml instance annotation: Exploiting label correlations with classifier chains. **Knowledge and Information Systems**, Springer, v. 43, n. 1, p. 53–79, 2015.

CHARTE, F. et al. A first approach to deal with imbalance in multi-label datasets. In: PAN, J.-S. et al. (Ed.). **Hybrid Artificial Intelligent Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 150–160. ISBN 978-3-642-40846-5.

CHARTE, F. et al. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. **Neurocomputing**, Elsevier, v. 163, p. 3–16, 2015.

EBERSBACH, M.; HERMS, R.; EIBL, M. Fusion methods for icd10 code classification of death certificates in multilingual corpora. In: **CLEF (Working Notes)**. [S.l.: s.n.], 2017.

FALEIROS, T. d. P. **Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais**. Tese (Doutorado) — Universidade de São Paulo, 2016.

FILHO, P. P. B.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013.

FONSECA, E. R.; ROSA, J. L. G. Mac-morpho revisited: Towards robust part-of-speech tagging. In: **Proceedings of the 9th Brazilian symposium in information and human language technology**. [S.l.: s.n.], 2013.

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

HERRERA, F. et al. Multilabel classification. In: **Multilabel Classification**. [S.l.]: Springer, 2016. p. 17–31.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. **science**, American Association for the Advancement of Science, v. 313, n. 5786, p. 504–507, 2006.

HO, K.-W. **Movies' Genres Classification by Synopsis**. [S.l.]: Citeseer, 2011.

HOANG, Q. Predicting movie genres based on plot summaries. **arXiv preprint arXiv:1801.04813**, 2018.

HOFFMAN, M.; BACH, F. R.; BLEI, D. M. Online learning for latent dirichlet allocation. In: **advances in neural information processing systems**. [S.l.: s.n.], 2010. p. 856–864.

HUANG, Y.-F.; WANG, S.-H. Movie genre classification using svm with audio and video features. In: SPRINGER. **International Conference on Active Media Technology**. [S.l.], 2012. p. 1–10.

IVASIC-KOS, M.; POBAR, M.; IPSIC, I. Automatic movie posters classification into genres. In: SPRINGER. **International Conference on ICT Innovations**. [S.l.], 2014. p. 319–328.

KIM, Y. et al. Character-aware neural language models. In: **Thirtieth AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2016.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **International conference on machine learning**. [S.l.: s.n.], 2014. p. 1188–1196.

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

PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900.

PETERS, M. E. et al. Deep contextualized word representations. In: **Proc. of NAACL**. [S.l.: s.n.], 2018.

QUINLAN, J. R. **C4. 5: programs for machine learning**. [S.l.]: Elsevier, 2014.

RAHMAN, R. I.; KADIR, S. et al. **Genre classification of movies using their synopsis**. Tese (Doutorado) — BRAC University, 2017.

RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. [S.l.]: Cambridge University Press, 2011.

RASHEED, Z.; SHEIKH, Y.; SHAH, M. On the use of computable features for film classification. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 15, n. 1, p. 52–64, 2005.

READ, J. et al. Classifier chains for multi-label classification. **Machine learning**, Springer, v. 85, n. 3, p. 333, 2011.

READ, J. et al. MEKA: A multi-label/multi-target extension to Weka. **Journal of Machine Learning Research**, v. 17, n. 21, p. 1–5, 2016. Disponível em: <http://jmlr.org/papers/v17/12-164.html>.

REHUREK, R.; SOJKA, P. Software framework for topic modelling with large corpora. In: CITESEER. **In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**. [S.l.], 2010.

SALAKHUTDINOV, R.; HINTON, G. Deep boltzmann machines. In: **Artificial intelligence and statistics**. [S.l.: s.n.], 2009. p. 448–455.

SUGANO, M. et al. Shot genre classification using compressed audio-visual features. In: IEEE. **Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)**. [S.l.], 2003. v. 2, p. II–17.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. **Journal of language and social psychology**, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining (IJDWM)**, IGI Global, v. 3, n. 3, p. 1–13, 2007.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Mining multi-label data. In: **Data mining and knowledge discovery handbook**. [S.l.]: Springer, 2009. p. 667–685.

TSOUMAKAS, G.; VLAHAVAS, I. Random k-labelsets: An ensemble method for multilabel classification. In: SPRINGER. **European conference on machine learning**. [S.l.], 2007. p. 406–417.

TURNER, M. D. et al. Automated annotation of functional imaging experiments via multi-label classification. **Frontiers in neuroscience**, Frontiers, v. 7, p. 240, 2013.

WEHRMANN, J.; BARROS, R. C. Convolutions through time for multi-label movie genre classification. In: ACM. **Proceedings of the Symposium on Applied Computing**. [S.l.], 2017. p. 114–119.

WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.

ZHOU, H. et al. Movie genre classification via scene categorization. In: ACM. **Proceedings of the 18th ACM international conference on Multimedia**. [S.l.], 2010. p. 747–750.

ZHOU, L. et al. A comparison of classification methods for predicting deception in computer-mediated communication. **Journal of Management Information Systems**, Taylor & Francis, v. 20, n. 4, p. 139–166, 2004.