Content-based and Knowledge-enriched Representations for Classification Across Modalities: A Survey

NIKIFOROS PITTARAS, National Kapodistrian University of Athens and NCSR “Demokritos”
GEORGE GIANNAKOPOULOS, NCSR “Demokritos”
PANAGIOTIS STAMATOPOULOS, National Kapodistrian University of Athens
VANGELIS KARKALETSIS, NCSR “Demokritos”

This survey documents representation approaches for classification across different modalities, from purely content-based methods to techniques utilizing external sources of structured knowledge. We present studies related to three paradigms used for representation, namely (a) low-level template-matching methods, (b) aggregation-based approaches, and (c) deep representation learning systems. We then describe existing resources of structure knowledge and elaborate on the need for enriching representations with such information. Approaches that utilize knowledge resources are presented next, organized with respect to how external information is exploited, i.e., (a) input enrichment and modification, (b) knowledge-based refinement and (c) end-to-end knowledge-aware systems. We subsequently provide a high-level discussion to summarize and compare strengths/weaknesses of the representation/enrichment paradigms proposed, and conclude the survey with an overview of relevant research findings and possible directions for future work.

CCS Concepts: • **Computing methodologies → Knowledge representation and reasoning**; **Natural language processing**; **Image representations;**

Additional Key Words and Phrases: Classification, enrichment, representations, semantics

ACM Reference format:
Nikiforos Pittaras, George Giannakopoulos, Panagiotis Stamatopoulos, and Vangelis Karkaletsis. 2023. Content-based and Knowledge-enriched Representations for Classification Across Modalities: A Survey. *ACM Comput. Surv.* 55, 14s, Article 312 (July 2023), 40 pages.
https://doi.org/10.1145/3583682

1 INTRODUCTION

A vast amount of diverse data are prevalent in our digital media ecosystem. Large quantities of text, images, audio, and video are created and circulated on the Internet, from journalism websites, blogs and academia-related content, to fiction literature portals, and social media. Efficient management, browsing, and consumption of this content depends on accurate discovery and search operations,

Authors’ addresses: N. Pittaras, National Kapodistrian University of Athens, Dept. of Informatics and Telecommunications, Athens 15784, Greece, NCSR “Demokritos”, Institute of Informatics and Telecommunications, Agia Paraskevi, Athens 15310, Greece; emails: npittaras@di.uoa.gr, pittarasnikif@iit.demokritos.gr; G. Giannakopoulos and V. Karkaletsis, NCSR “Demokritos”, Institute of Informatics and Telecommunications, Agia Paraskevi, Athens 15310, Greece; emails: {ggianna, vangelis}@iit.demokritos.gr; P. Stamatopoulos, National Kapodistrian University of Athens, Dept. of Informatics and Telecommunications, Athens 15784, Greece; email: takis@di.uoa.gr.

0360-0300/2023/07-ART312 $15.00
https://doi.org/10.1145/3583682

applied on massive data collections. To this end, developing robust classification methods for tagging and categorization has been crucial in the era of big data. Classification systems [Aggarwal 2015] automatically assign labels to new instances, facilitating efficient organization and categorization of large volumes of data with little to no human involvement. They are applied to a wide array of commercial, industrial and artistic applications, from phishing and spam detection [Subramaniam et al. 2010; Whittaker et al. 2010], medical imaging, style transfer, and optical character recognition [Jing et al. 2019; Razzak et al. 2018; Singh 2013], to speaker diarization and genre classification [Anguera et al. 2012; Sturm 2012]. The ubiquitous adoption of such models enables the automation of such tasks, avoiding the comparatively prohibitive cost of manual human effort on such a large scale.

At the same time with the rapid growth of available content, a systematic collection, structured formatting, and storage of knowledge has accompanied the growth of artificial intelligence research and the development of commercial AI-powered solutions. As a result, a wealth of curated, high quality information is available and applicable in classification systems and **machine learning** (**ML**) tasks, ranging from fine-grained linguistic and audiovisual information to high-level conceptual ontologies [Deng et al. 2009; Ganitkevitch et al. 2013; Gemmeke et al. 2017; Miller 1995]. However, the utilization of such resources for broad-range solutions has been lacking.

An important classification component is building the data representation, i.e., mapping real-world objects fed to the system (e.g., e-mails, photographs, recordings) into a feature collection that can be processed by the ML system [Storcheus et al. 2015]. Since this mapping is often the sole information source for subsequent components, producing high-quality representations is a crucial part for efficient classification. This often requires the construction of representations that go beyond low-level local pattern-matching (e.g., token/ngram frequencies in text, local template matching in audiovisual content), but encapsulate complex, conceptual information [Bengio 2009, 2011]. A lot of research effort has pursued this via handcrafted feature engineering/transformations, as well as automated methods for content-based representation learning [Bengio et al. 2013; Zheng and Casari 2018]. However, engineered approaches tend to rely on empirical expert knowledge specific to the modality, data, domain and/or task at hand and is often based on rigid heuristics (e.g., text token bag hyperparameters, visual descriptor templates or audio signal/temporal/segmentation-based measures). On the other hand, representation learning methods operate on vast amounts of data, require considerable computational resources and energy [Strubell et al. 2019] and heavily rely on the distributional hypothesis [Harris 1954] to arrive at sets of features from the bottom-up, that hopefully encapsulate useful semantics.

In light of these limitations, this article investigates the utilization of high-level information (e.g., conceptual, semantic, relational) encoded in knowledge resources in the classification pipeline, with a focus on enriching data representations fed to the predictive algorithm. We provide a brief survey of representation methods used for classification for different modalities, along with representation enrichment approaches from external information sources.

1.1 The Structure of a Classification Problem

Classification entails applying a meaningful category (i.e., a semantic label/class) to a piece of data. In ML, a typical classification problem consists of the following components [Sebastiani 2002].

(1) Inputs: A labeled dataset $D : (d_i, L_i)$, $i = [1, 2, \ldots, N]$, $L_i \in L$ where d_i is an input instance, L is the set of all available dataset labels and L_i are the labels associated with the $i - th$ instance. A label $l \in L$ is a semantic tag related to the content of d_i, either directly (e.g., topic or sentiment classification) or indirectly (e.g., related to content generation, such as in authorship attribution and style classification tasks). Classification problems can be

characterized by the number of instance candidate labels $|L|$ – e.g., binary (two available labels, that may map to a "yes" or "no" answer, e.g., hate speech detection, medical image disease prediction, speech pathology detection) and multiclass (more than two candidate labels, e.g., as in text topic classification, visual object recognition, and music genre classification). The maximum number of label annotations per instance $M = \underset{i=1 \ldots N}{max} |L_i|$ renders the task as a single-label ($M = 1$, i.e., one label allowed per instance – e.g., sentiment analysis, handwritten digit recognition and speaker diarization) or a multi-label ($M > 1$, i.e., multiple possible labels per instance – e.g., document topic classification, visual object recognition, and music genre classification).

(2) Preprocessing: preliminary data operations [Camastra and Vinciarelli 2015; García et al. 2015], e.g.,:
- Data Augmentation: when dataset-related limitations impact performance (e.g., sample scarcity, label imbalance) operations that modify the input collection may be employed, such as data augmentation and under/oversampling [He and Ma 2013; Shorten and Khoshgoftaar 2019; Van Dyk and Meng 2001].
- Data cleaning [Chu et al. 2016] – e.g., handling undesirable input patterns (e.g., non-alphanumerics/whitespace in text, image/audio denoising, frequency filtering, and so on. [Bhattacharyya 2011; Burges et al. 2002])
- Filtering – e.g., stopword handling, stemming, lemmatization, POS extraction in text [Jivani 2011; Ratnaparkhi 1996; Silva and Ribeiro 2003], normalization/equalization in images/audio [Bai and Chen 2007; Bhattacharyya 2011; Pei and Lin 1995] as well as modality-shifting operations [Knight et al. 2020].
- Segmentation – e.g., word/sentence splitting and tokenization in text [Vijayarani et al. 2015], region/color segmentation in images [Bhattacharyya 2011], temporal/spectral segmentation and source separation in audio [Chang et al. 2021; Mesgarani et al. 2006; Theodorou et al. 2014; Zaitoun and Aqel 2015]

(3) Representation: A representation mapping [Martinčić-Ipšić et al. 2019] converts preprocessed inputs into a suitable format, with respect to computational costs of subsequent processing and efficiency of learning [Bengio 2009]. This is often realized through mappings to a vector space [Salton et al. 1975] in which objects are represented by ordered sets of attributes deemed useful for the task by algorithms or human engineers. Although different approaches have been explored [Giannakopoulos et al. 2012; Jolion and Kropatsch 2012; Sonawane and Kulkarni 2014], vector-based representations of data are both straightforward and allow exploiting the rich tradition, advances and tools of vector algebra and calculus, as other disciplines have done before (e.g., mathematics, physics and multiple engineering fields); as a result, vector formats are either required by or compatible with the vast majority of data analysis and ML approaches today.

(4) Classification: finally, a learning algorithm is trained to solve the classification task: given a representation x_i of the input instance $d_i \in D$, produce an optimal approximation of L_i with respect to a performance measure, ensuring good generalization ability on unseen data [Mohri et al. 2018; Sokolova and Lapalme 2009].

With this framework in mind, we will explore knowledge utilization methods for improving classification tasks.

1.2 Injecting Knowledge into Classification

Multiple sources of structured human knowledge exist, that can be utilized in classification applications. This article explores avenues of introducing such information in representations for the classification task, towards improving categorization performance, as these have been discussed in the existing literature. Our specific contributions include:

- A novel, complexity-focused perspective on representation construction paradigms, paired with a large body of indicative literature review of related approaches using text, image and audio data, presented in Section 2.
- A novel categorization of representation enrichment methods from a perspective knowledge utilization mechanisms, along with multiple representative literature over three modalities, provided in Section 4.
- A detailed review of knowledge resources used for enrichment methods discussed, presented in Section 3.
- A comparative, high-level view of proposed content-based and enrichment paradigms with respect to multiple qualitative axes of representation desiderata, presented in Section 5.
- A critical overview of the totality of covered work, along with the body of findings and suggested future directions in representation enrichment regardless of the underlying modality, provided in Section 6.

There have been multiple surveys focusing on modality-oriented classification [Fu et al. 2010; Lu and Weng 2007; Minaee et al. 2021; Sebastiani 2002] as well as reviews on knowledge integration for ML tasks. The study in Altınel and Ganiz [2018] focuses on quantitative comparisons of experimental results over broad approaches for text classification. The work of Turney and Pantel [2010] investigates semantic variants of vector space models [Salton et al. 1975] and their utilization in text ML tasks. In Camacho-Collados and Pilehvar [2018], the authors focus on sense representations of text in both supervised and unsupervised settings, along with application and evaluation approaches. The survey in Ferrone and Zanzotto [2020] covers symbolic, distributed and distributional features for NLP tasks, while [Borghesi et al. 2020] provides a brief overview on augmenting deep learning with external knowledge expressed in constraints.

This study complements existing surveys by approaching knowledge enrichment with emphasis on:

- The classification task; our examination and the representations covered are utilized for classification
- A holistic review of representation extraction; our primary focus is representations construction methods, from simple low-level feature extraction to deep representation learning [Bengio et al. 2013].
- Different approaches of knowledge infusion to representations; we examine a broad range of representation enrichment methods, from low-level semantic features to end-to-end representation learning modifications.
- Different knowledge resources; we cover multiple kinds of information and knowledge repositories (e.g., graphs, lexicons, collection of name-value information, linked data)
- Different modalities; we examine the aforementioned axes on studies dealing with text, image, and audio data

1.3 Article Structure

The article is structured as follows: we begin with content-based (knowledge-agnostic) representations for classification over different modalities in Section 2. In Section 3 we expand on the motivation for knowledge utilization and present a collection of usable resources. Section 4 covers representation enrichment via knowledge exploitation, for classification tasks on different modalities. A qualitative comparison over the proposed paradigms is facilitated in Section 5. We discuss findings and outcomes in Section 6 and close by presenting conclusions and future work avenues in Section 7.

2 KNOWLEDGE-AGNOSTIC REPRESENTATIONS FOR CLASSIFICATION

In this section, we discuss existing data representation approaches for classification tasks that do not explicitly consider external information sources. We structure the discussion by organizing studies into 3 categories/paradigms:

- **Low-level and template-matching (LLTM) methods:** We begin with approaches that rely on matching predefined templates on the input data. The collection of responses of the template matching constitutes the representation output and a corresponding vector space embedding. Such approaches are covered in Section 2.1, and generally correspond to simple, engineered and isolated components that perform intuitive measurements on real-world objects, mapping them to a computationally usable numeric representation to be fed down the pipeline. Examples include, e.g., bags of words and features, audio and visual descriptor templates, statistical/signal measures of data streams, and so on. [Boyer et al. 1999; Jörgensen et al. 2001; Sebastiani 2002].
- **Aggregation-based (AGB) methods:** here we include approaches that group, transform and/or combine low-level representations from the previous category into higher level features, simultaneously facilitating improvements in terms of computational efficiency, redundancy filtering, and dimensionality reduction. Works in this category showcase initial representation and ML research efforts towards improving representation quality via content-based means, using engineered and pre-defined aggregation and/or post-processing operations. Approaches may include clustering, topic modeling, and decomposition methods [Miettinen 2009; Rokach and Maimon 2005; Uys et al. 2008]. They are covered in Section 2.2.
- **Deep representation learning (DRL) methods:** the final group covers approaches that heavily rely on a hierarchy of modular, nonlinear components for representation learning. We focus on **neural network** (**NN**) models and deep learning, that can produce very high-level information, i.e., rich features that correlate to high-level abstract/conceptual/context-aware information. This category reflects recent deep learning trends that discourage the explicit definition and over-engineering of the feature extraction procedure. Instead, arriving at richer and more efficient representations is entrusted to neural systems, which are set up to automatically discover useful features by learning from big data. These methods include convolutional, recurrent and transformer NNs [Gu et al. 2018; Medsker and Jain 2001; Tay et al. 2023] and are presented in Section 2.3.

The proposed grouping above enables the identification of the general feature generation approach adopted for each work. Upcoming sections focus on each category, presenting related work that adopts such approaches for classification of different modalities. Notably, the saliency and holistic nature of the proposed content-based paradigms render them relevant to a very large number of different classification tasks and domains; to this end, works covered for each category in the next sections should not be viewed as an exhaustive enumeration of all relevant approaches, but rather as indicative characteristic, influential and/or recent examples which illuminate the overall approach and methods typically pursued in the paradigm. Classification pipeline details (e.g., features, classifiers, metrics) for each study in the content-based literature covered is presented in Table 1.

2.1 Template Matching and Low-level Approaches

2.1.1 Overview. In this section, we examine LLTM: representations that rely on matching templates on the input data [Bengio 2009], producing low-level responses. Given a representation template, we can discriminate between "local" (the template is applied on an input subsection) and

Table 1. Indicative Studies using Knowledge-agnostic Representations

citation	mod.	category	representation	labeling	classifiers	metrics
[Badawi and Altınçay 2014]	TXT	LLTM	BOW, termsets	BIN	k-NN	F1
[Trstenjak et al. 2014]	TXT	LLTM	TFIDF	MC-SL	k-NN	ACC
[Zhang and Wu 2015]	TXT	LLTM	BOW, extension	MC-SL	NB	P, R, F1
[Sowmya et al. 2016]	TXT	LLTM	TFIDF	MC-ML	k-NN, Rocchio	mAP
[Thirumoorthy and Muneeswaran 2021]	TXT	LLTM	BoW	MC-SL	SVM, NB	P, R, F1, ACC
[Zhang et al. 2011]	TXT	AGB	TFIDF, LSA	MC-SL	SVM	P, R
[Giannakopoulos et al. 2012]	TXT	AGB	NGG	MC-SL	graph similarity	P
[Zareapoor and Seeja 2015]	TXT	AGB	BoW, PCA, LSA	MC-SL	RF	AUC, ACC
[Ye et al. 2017]	TXT	AGB	TFIDF, LDA	MC-SL	SVM	P, R, F1
[Škrlj et al. 2021]	TXT	AGB	TFIDF, Word2Vec, Evolution	MC-SL	SVM, NEURAL, LR	ACC
[Liu et al. 2015]	TXT	DRL	SKIPGRAM, LDA	MC, SL	LINEAR	PR, R, F1, ACC
[Yang et al. 2018]	TXT	DRL	SKIPGRAM	BIN	NEURAL	ACC
[Sun et al. 2019a]	TXT	DRL	TRANSFORMER	BIN / MC, SL	NEURAL	ACC
[Chen et al. 2020]	TXT	DRL	TFIDF, CNN	MC, SL	NEURAL	ACC
[Pan et al. 2022]	TXT	DRL	TRANSFORMER	MC, SL	NEURAL	ACC
[Risojević et al. 2011]	IMG	LLTM	GIST, Gabor	MC-SL	SVM	ACC
[Amato et al. 2015]	IMG	LLTM	SIFT, SURF, ORB, BRISK	MC-SL	k-NN, similarity	ACC, F1
[Bian et al. 2017]	IMG	LLTM	SIFT, LBP	MC-SL	KELM	ACC
[Prasad and Mary 2019]	IMG	LLTM	HoG, BRISK, LBP, KAZE	MC-SL	SVM	ACC
[Ningtyas et al. 2022]	IMG	LLTM	LBP, GLCM	MC-SL	k-NN	ACC
[Zou et al. 2016]	IMG	AGB	SIFT, LBP, SPM, LLC, KMeans	MC-SL	KCR	ACC
[Ilea et al. 2016]	IMG	AGB	RCovD, GMM, FV	MC-SL	kSVM	ACC
[Srivastava et al. 2019]	IMG	AGB	LBP, CFC	MC-SL	SVM	P, R, S, F1, ACC
[Zhu et al. 2019]	IMG	AGB	SURF, BRISK, KMeans	MC-SL	SVM	ACC
[Bodine and Hochbaum 2022]	IMG	AGB	pixel, PCA	Binary, MC-SL	DT	ACC
[Huang et al. 2017]	IMG	DRL	CNN, RESIDUAL	MC, SL	NEURAL	ACC
[Dosovitskiy et al. 2021]	IMG	DRL	TRANSFORMER, CNN	MC, SL	NEURAL	ACC
[Touvron et al. 2022]	IMG	DRL	MLP, RESIDUAL	MC, SL	MLP	ACC
[Liu et al. 2021]	IMG	DRL	TRANSFORMER	MC, SL	NEURAL	ACC
[Chen et al. 2021]	IMG	DRL	TRANSFORMER	MC, SL	NEURAL	ACC
[Laurier et al. 2009]	AU	LLTM	signal, musical, psych.	MC-SL	kSVM	ACC
[Valero and Alias 2012]	AU	LLTM	spectral, MFCC, MPEG7	MC-SL	DT, SVM, MLP, k-NN	ACC
[Maršík et al. 2014]	AU	LLTM	signal, spectral, musical	MC-SL	MLP	P, R
[Zahid et al. 2015]	AU	LLTM	signal, spectral, temporal, MFCC	MC-SL	SVM, MLP, rule-based	ACC
[Meister et al. 2022]	AU	LLTM	signal, spectral, temporal, MFCC	MC-SL	SVM, RF, LR, k-NN, ADAboost	ACC, AUC
[Lee and Ellis 2010]	AU	AGB	MFCC, GMM, PLSA	MC-SL	kSVM	AP
[Kim et al. 2012]	AU	AGB	MFCC, VQ, LDA	MC-SL	kSVM	F1
[Grosse et al. 2007]	AU	AGB	MFCC, SISC, spectral	MC-SL	GDA, SVM	ACC
[Baniya et al. 2014]	AU	AGB	MFCC, signal, spectral, musical, PCA, MRMR	MC-SL	SVM	ACC
[Zeghidour et al. 2021]	AU	AGB	Gabor, GMM	MC-SL, ML	MLP	ACC, AUC
[Choi et al. 2016]	AU	DRL	CNN	MC, SL	NEURAL	ACC
[Hershey et al. 2017]	AU	DRL	CNN, INCEPTION, RESIDUAL	ML	NEURAL	mAP, AUC
[Nanni et al. 2021]	AU	DRL	spectral, CNN	MC, SL	NEURAL	mAP
[Gong et al. 2021]	AU	DRL	spectral, CNN	MC, SL	NEURAL	ACC
[Wang and Oord 2021]	AU	DRL	spectral, CNN, RESIDUAL	MC, SL	NEURAL	ACC, mAP

LLTM, AGB and DRL refer to the categories outlined in Section 2. MC, SL, and ML labeling refers to multiclass, single-label and multi-label configurations respectively. Evaluation measures ACC, PR, RE, F1, AUC, and (m)AP refer to accuracy, precision, recall, F1-measure, area under curve and (mean) average precision, respectively. Entries in the representation/classifier columns refer to acronyms described in the text, or intuitive categories of algorithms/approaches. NEURAL refers to using an appropriate layer for classification in the network output (e.g., dense followed by softmax).

global (the template spans the entire input instance) applications of the extraction process [Mikolajczyk et al. 2005; Zhang et al. 2007]. Local templates may regard the input as key-value pairs: keys locate distinct attributes (e.g., an individual word/ngram in text, a region/point of interest in images and audio) while values correspond to a magnitude of match or *weight* of the template in that location. Terms may be easily delineable in the source data (e.g., individual words in text, detected keypoints in images, specific temporal slice/peak in audio) or lack high-level semantics and an intuitive explanation—the latter can affect the interpretability of the representation and, down the line, of the entire classification pipeline [Danilevsky et al. 2020; Došilović et al. 2018]. Global features usually provide coarse information on a narrow view of the input—e.g., color distribution information in images, sentiment/grammaticality scores in text, SNR values for audio, and so on. [Oliva and Torralba 2006].

2.1.2 *Approaches.* A popular representation is a **Vector Space model** (**VSM**) [Salton et al. 1975], which projects data into a vector $v \in \mathbb{R}^d$ that can be manipulated with distance measures and Linear algebra constructs [Basseville 1989; Cha 2007] to process/compare instances. The **Bag**

of Words/Features (**BoW/BoF**) [Salton and Buckley 1988; Sebastiani 2002] is a popular VSM that produces count-based weights for points/regions of interest in the input. BoF is a popular baseline for text, where semantically salient terms are easily identifiable and delineated by syntax and grammar. Common weighting schemes include boolean, term, and document frequency (BF, TF, DF), denoting presence, instance, and collection-level counts of single or n-tuples of terms (n-grams) in the text. The work in Badawi and Altınçay [2014] utilizes BoW features, along with an investigation on term weighting and selection in the binary classification of articles and biomedical documents. "Termset features" are introduced—i.e., tuples of document terms (e.g., words) where the feature activates if either one or both terms are detected. In Zhang and Wu [2015], the authors reduce sparsity by building an extension library, using bigram conditional probabilities in the text sequence and word-to-category similarity, based on word counts. Given a text, additional features are inserted with respect to their similarity score to the original feature set and introduced threshold heuristics. The extended feature set is then used to build the BoF representation.

The Term frequency-inverse document frequency (**TFIDF**) weighting scheme [Salton and Buckley 1988] normalizes term counts by DF weight, reducing the importance of tokens that occur too often in the document collection and behave like stopwords. In Trstenjak et al. [2014], the authors build a term-document frequency matrix to map articles to word TFIDF vectors, applying log-scaling and renormalization to improve robustness to varying document lengths. The authors in Sowmya et al. [2016] apply TFIDF to Wikipedia articles; document weights are pooled to category-level counts in order to build "centroid" vectors for each class, subsequently normalized by intra-class DF scores. In Thirumoorthy and Muneeswaran [2021], the authors propose feature selection schemes with respect of term and document frequency scores within and across classes. They use an evolutionary method, rich/poor population-based methods [Moosavi and Bardsiri 2019], and a classification-based fitness objective to optimize the final feature subset.

Contrary to text, BoF application in the visual domain is not straightforward: images lack clear semantic boundaries ("visual terms" are hard to delineate) and pixel-level approaches are intractable for most real-world tasks. Thus, visual LLTMs employ methods that apply (a) *detection*, i.e.. locate regions of interest in the image [Tuytelaars and Mikolajczyk 2008], and (b) *description*, which applies low-level templates for building a representation for each such detected region.

A popular method is **Scale Invariant Feature Transform** (**SIFT**) [Lowe 2004; Younes et al. 2012]. It describes keypoints by normalized histograms of gradient orientations in the pixel intensities of a co-centric patch, that are largely invariant to shifts in illumination, viewpoint, rotation, and scale. SIFT is adopted for a variety of recognition tasks, e.g., detection/description in [Amato et al. 2015] for a local feature-based landmark classification along with other descriptors [Bay et al. 2008], while a similar approach extracts SIFT from a dense regular grid with patch overlaps [Bian et al. 2017].

Other methods extract fine-grained information suitable for texture, such as Gabor features [Manjunath and Ma 1996; Mehrotra et al. 1992], that involve the application of Gabor filterbanks. Gabor filters are applied on separate color channels in Risojević et al. [2011], using mean and stdev values over multiple scales and orientations, along with spatial envelope GIST features [Oliva and Torralba 2001]. Another approach is **Local Binary Patterns** (**LBP**) [Ojala et al. 2002], that extracts fine-grained, rotation-invariant binary-value histograms, via simple pixel-level comparisons between a center and its radial neighbors. LBP has been evaluated with different tasks, global/local contexts, resolutions, and focal configurations [Bian et al. 2017; Prasad and Mary 2019]. The authors in Ningtyas et al. [2022] utilize LBP along with **Gray Level Co-occurence Matrix features** (**GLCM**) to capture texture information for leaf classification.

ORB [Rublee et al. 2011] improves upon BRIEF [Calonder et al. 2010] features, adding "steering" mechanisms for rotation invariance and noise resistance, and is used in tasks like monument classification [Amato et al. 2015]. Further methods include **Histograms of gradients** (**HoG**) [Dalal and Triggs 2005], which computes distributions of pixel intensity gradient orientations, BRISK [Leutenegger et al. 2011], producing pairwise intensity comparison as binary features and KAZE [Alcantarilla et al. 2012], which applies nonlinear diffusion filtering for smoothing and multiscale operation; these are used for feature extractor/detection in various tasks [Amato et al. 2015; Prasad and Mary 2019].

In audio data, similar semantic ambiguity exists, as in the visual domain; However, given its one-dimensional and temporal structure, LLTM methods often apply simple features within time-segmented frames or in a global context.

A popular method is to capture statistical signal properties (e.g., mean, variance, extrema) and utilize the responses as feature vectors; for example, in Maršík et al. [2014] the authors consider RMS amplitudes as audio volume estimates, along with a self-similarity computation that counts the similar segments in a music piece. Further, in Zahid et al. [2015], signal sign change (zero crossing rate, ZCR) averages, short-time signal energy and periodicity analysis features are concatenated across audio window frames and used toward capturing repeated acoustic patterns.

Another approach switches to the frequency domain to mine low-level features from audio spectra. This is often achieved via **Fast Fourier Transform** (**FFT**) [Bracewell and Bracewell 1986]; FFT maps time-domain signals to frequency spectrograms via Fourier Analysis on small overlapping time windows. This is used in works like [Laurier et al. 2009], where the authors extract spectral statistic estimates like kurtosis, skewness, flatness, and flux for music emotion classification. **Mel-frequency cepstral coefficients** (**MFCC**) [Slaney 1998] involve multiple transformation, scaling, and normalization steps for short-term power spectrum description of an audio signal. It is used in studies like [Maršík et al. 2014], using mean and covariance statistics and [Zahid et al. 2015], along with spectral flux scores. Furthermore, **Gammatone Cepstral Coefficients** (**GTCC**) is a biologically-inspired modification to MFCC based on Gammatone filter functions with equivalent rectangular bandwidth bands; it is proposed and utilized in Valero and Alias [2012].

Some features target musicality/rythm-based information by using frequency binning and temporal progression statistics, or model psycho-acoustic phenomena by considering the operation of the human hearing system. For instance, timbral, perceptual and tonal information via dissonance and loudness measures are extracted in [Laurier et al. 2009], along with "danceability", **beats per minute** (**BPM**), ZCR, and chord change features. The authors in Maršík et al. [2014] use musical LLTM like BPM, probabilistic estimates of chord root transitions, and musical keys. Furthermore, the work in Meister et al. [2022] uses an engineered feature bank composed of temporal, spectral, cepstral and tonal responses, which are ranked and evaluated for COVID patient classification.

In summary, this section showcased approaches that utilize LLTM features for classification; we now move on to methods that transform, manipulate and aggregate this information towards improving performance and tractability.

2.2 Aggregation-based Methods

2.2.1 Overview. In this section, we explore approaches that rely on aggregating, combining and/or transforming lower-level representations to arrive at higher-level features, with respect to the abstractness and richness of the information encapsulated. Aggregation methods produce mid-level representations from low-level inputs, using engineered and/or learned functions rather than directly exploiting token-based statistics. In the scope of our analysis, these techniques correspond to the first attempts of improving the semantic content and richness of low-level representations by applying aggregation as a post-processing component. Contrary to low-level features, AGB methods usually build distributed representations [Hinton et al. 1984; Rumelhart 1986]: i.e., resulting semantics are spread or "distributed" over multiple dimensions in the embedding space [Bengio 2009], arriving at compact, robust features but sacrificing explainability. Dimensionality reduction is often facilitated by these approaches, mitigating the problem of the curse of dimensionality [Bellman 2013] while simultaneously maintaining or improving the expressive power of the output feature set.

2.2.2 Approaches. Aggregation methods have presented different avenues for fusing LLTM features. Modality-specific delineation of semantically meaningful terms play an important role to the usefulness of different approaches.

For text data, local LLTM methods deal with distinct words, characters, and ngrams; as a result, the generated vocabulary can quickly scale to very large sizes, reducing system performance and accuracy [Bengio et al. 2005]. Thus, aggregation methods for text will often aim at shifting the representation from the vocabulary space to one more dense and compact, and simultaneously preserves the majority of the information content of its original counterpart.

A popular class of AGB methods is matrix decomposition techniques [Golub 1969], where a $M \times d$ matrix of M, d-dimensional features are converted to a $M \times k$ matrix, $k \leq d$. For instance, a popular method is **Latent Semantic Analysis** (**LSA**) [Deerwester et al. 1990], which produces document mappings to a set of k latent concepts. This is achieved by truncating the Singular Value Decomposition factors [Trefethen and Bau III 1997] of a term-document matrix, keeping the vectors with the k largest singular values. LSA is used in Zhang et al. [2011] to post-process TFIDF features and a multi-word term vector approach [Justeson and Katz 1995] for news articles categorization with linear SVMs. Additionally, **Principal Component Analysis** (**PCA**) [Jolliffe 2011] applies a change of base to the principal components of the original collection, that correspond to the covariance matrix of the original instances. Keeping a subset of k instances retains an optimal tradeoff with respect to variance retention and dimensionality reduction. PCA has been used in early works, for example in Li and Jain [1998], where term count features are post-processed with PCA and hierarchical clustering for news categorization with multiple classifiers. Additionally, in Zareapoor and Seeja [2015] both PCA and LSA are investigated for compression, reduction of complexity, and processing time of BoW features. Further, **Latent Dirichlet Allocation** (**LDA**) [Blei et al. 2003] generates topic models over Dirichlet priors, assuming distributions from instances to topics and from a topic to words. In Ye et al. [2017], the authors use TFIDF to model outlier words that are overlooked by LDA and include them for sentiment classification.

Other works adopt evolutionary methods to modify the representation; in Škrlj et al. [2021], weights for given diverse sets of input space features undergo mutation and crossover, while fitness is estimated by aggregate classification performance with multiple learners trained with SGD. Further, n-gram graph aggregation methods in Giannakopoulos et al. [2012] map documents to graphs, subsequently merged into class-level constructs. These are then compared with document-level graphs for categorization of the latter using different graph similarity measures.

In images, LLTM features generally produce non-scalar information, yielding large volumes of highly local responses. Many approaches use vector quantization [Gersho and Gray 1992] to fuse descriptor vectors and reduce workload to subsequent classification components, redundancy, and noise. Clustering methods are popular for this task; for example, Kmeans [Jain and Dubes 1988] merges populations of N local features to arrive at a pre-determined size of $k \ll d$ visual clusters. These serve as an artificial vocabulary for a visual BoW analog, with distance-based assignments of local features to "visual words". The bag vector is then used as a global image feature. KMeans vocabularies are used, e.g., in Zhu et al. [2019], over different combinations of SURF and BRISK for detection and description.

Modifications to the visual BoW include **Spatial Pyramid Matching** (**SPM**) [Lazebnik et al. 2006], which partitions and separately groups keypoints to preset image subdivisions. Further,

locality-constrained linear coding (**LLC**) [Wang et al. 2010] enforces locality constraints on KMeans via regularization and modification the membership procedure, making instances being supported by multiple codebook bases for reduced reconstruction error and sparsity. LLC is evaluated by the work in Zou et al. [2016], aggregating dense SIFT descriptors in combination with global LBP features. Transformation-oriented methods like PCA are applied in the visual domain in Bodine and Hochbaum [2022], where a modification of Decision Trees [Lewis and Ringuette 1994] is proposed, which uses a one-dimensional, single-feature Maximum Cut split criterion, in conjunction with two localized PCA-based methods for transforming and mapping decision features. The **vectors of locally-aggregated descriptors** (**VLAD**) approach [Delhumeau et al. 2013] replace binary cluster assignments with concatenations of accumulated subtraction residuals. Another modification is **Clustering with Fixed Centers** (**CFC**) [Srivastava et al. 2019], over LBP features and SURF keypoints. Descriptors are grouped into bags assuming a fixed cluster center with respect to response maxima per LBP histogram bin, with combined category-level bags being used as the global feature. Further, fisher vector encoding [Perronnin and Dance 2007] uses visual vocabularies built with **Gaussian Mixture Models** (**GMM**) [Reynolds 2009], using the log-likelihood GMM gradients under the data are as encoded codebook features. A generalization to Riemannian spaces is proposed in Ilea et al. [2016], generated by **Riemmanian Gaussian Mixture** (**RGM**) models. The method uses **region covariance descriptors** (**RCovD**) as input features, built by covariance information from on sliding image patches.

In the audio domain, AGB approaches may exploit sequential signal interdependencies to pool together feature responses spatially/spectrally close to each other, along aforementioned methods of segmentation into "acoustic words".

Aggregation via generative/probabilistic models has been used for classification, like GMMs in Lee and Ellis [2010]. There, MFCC inputs are combined with single and multiple-gaussian GMMs—pLSA is then applied on the built acoustic vocabulary. In the study of Kim et al. [2012] "Acoustic words" are formed from MFCC features via the LBG-VQ quantizer [Gersho and Gray 1992]. LDA is subsequently employed to generate audio-related topics over the audio vocabulary, with resulting weights serving as latent vectors for classification.

Like other modalities, matrix decomposition has also been applied for audio. In Baniya et al. [2014], the authors use PCA to reduce a feature set of spectral, dynamic, harmonic, and rhythm characteristics, along with higher order moments. **Sparse coding** (**SC**) [Lee et al. 2007] introduces sparsity bias to the encoding objective, in the form of a weighted L1 regularization term on the encoding vectors. A shift-invariant modification [Olshausen and Field 1996] proposed in Grosse et al. [2007] reconstructs the input signal to basis functions in all possible shifts to build temporally-invariant encodings. The authors apply computational efficiency improvements and compare the encodings with MFCC and raw spectrogram features. The approach in Zeghidour et al. [2021] applies a pipeline of filtering, pooling and compression/normalization composed of learnable steps, including Normalized Gabor convolutions, Gaussian low-pass filters and per-Channel Energy normalization [Wang et al. 2017], fed to MLPs for different categorization scenarios.

Having examined aggregation engineering of low-level features, the next section concludes the examination of knowledge-agnostic approaches by considering automatic DRL techniques.

2.3 Deep Representation Learning

2.3.1 Overview. Approaches covered so far used preconfigured templates locally interpolated on training data points, as well as manipulations of their responses in fixed, preconfigured steps. In this section, we focus on "deep" feature extractors, that aim at automatically learn multiple useful *feature hierarchies* from training data [Bengio 2009].

Typical representatives of this paradigm are graph-based computational models, such as the biologically inspired artificial **neural networks** (**NNs**) [Hubel and Wiesel 1962]. While *deep* generally refers to models that learn at semantically abstract, meaningful features, in the context of NNs, it can also refer to the size of the graph that implements the computation. Deep NNs are hierarchical models composed by multiple steps of nonlinear, learnable feature transformations. Contrary to previous paradigms, the line between features and classifiers is blurred: feature transformation, representation learning, and classification are simultaneously optimized in a holistic manner, learning the conversion of input data to prediction scores in an automatic, end-to-end fashion. This enables the design of efficient representation learners with fewer hard-coded, performance-critical parameters. As a result, deep approaches reflect further research efforts for improving representation semantics, diverging from rigorous feature and/or aggregation engineering and relying on a data-driven, automatic discovery of expressive features from scratch instead.

2.3.2 Approaches. DRL methods generally exploit unsupervised pretraining [Bengio et al. 2007], fine-tuning, and transfer learning [Zhuang et al. 2020], with fitted model weights used as initializations for supervised training/fine-tuning for classification, or used directly as standalone features [Bengio et al. 2013; Pittaras et al. 2017].

Neural language models (**NNLMs**) [Arisoy et al. 2012] are popular approaches in text, applying distributional learning over big data to learning language structure and feature hierarchies via statistical learning of token distributions. NNLM-based pretraining and transfer learning DRL are an effective approach for deep feature extraction, as comparisons between NNLM-based methods, LLTM, and AGB techniques have indicated [Baroni et al. 2014],

Word embeddings are a popular example—for instance, Word2Vec [Mikolov et al. 2013] is pretrained via predicting center words given a context (CBOW), or the context, given the center word (Skipgram). Multiple studies exploit the effectiveness of transfer learning with Word2Vec features. In Yang et al. [2018], Word2Vec is pretrained over social media data with concatenated word vectors feeding a **Convolutional Neural Network** (**CNN**) with pooling components. Skipgram embeddings initialize the system in Liu et al. [2015], where LDA is used to build word and topic embeddings, exploring transfer learning variants to embed topics, word-topic pairs, or to concatenate topic and word vectors.

Some approaches utilize count-based inputs/seeds for faster operation and improved compatibility. In Chen et al. [2020], input TFIDF features are transformed into 2-dimensional matrices, fed to a CNN architecture equipped with pooling and fully-connected components. Other methods use transformers [Vaswani et al. 2017], relying on attention mechanisms [Bahdanau et al. 2015] to facilitate sequence modeling and learning, instead of recurrence or convolution. These methods have been shown to produce powerful, transferable representations, explored in studies like [Sun et al. 2019a], where the authors propose ways of fine-tuning BERT [Devlin et al. 2019] for efficient classification. Enhancements include additional pre-training, single/multi-task fine-tuning, and varying learning rates per transformer layer. In Pan et al. [2022], the authors introduce data augmentation by generating adversarial examples with FGSM [Goodfellow et al. 2015] by perturbing the word embedding matrix of a transformer, using the result for contrastive learning of noise invariant representations and improved performance on multiple NLU tasks including sentiment analysis.

In images, the large semantic gap between pixel-level and conceptual information has encouraged the design of DRL approaches. CNN architectures have been a very popular in this domain, with early successful approaches contributing to the rise in popularity of deep learning [Krizhevsky et al. 2012]. CNNs typically repeat layers of convolution, pooling and normalization, followed by fully-connected (dense) layers with dropout [Srivastava et al. 2014] and softmax normalization, and have been popular choices for large-scale image classification tasks [Russakovsky et al. 2015].

Different CNN topologies have been proposed to allow efficient training of larger models, e.g., residual connections, which preserve signals via additive identity operations and inception modules, which propose blocks of dimensionality reduction, pooling filtering and concatenation to improve scalability [He et al. 2016; Szegedy et al. 2015]. Additional approaches include the "densely connected" architecture, where layer outputs are linked to the inputs of all subsequent layers and vice versa, and residual connections replaced with feature concatenation. This topology is evaluated using the DenseNet model [Huang et al. 2017] on a wide range of image classification tasks. The model requires fewer parameters than ResNets, achieved by utilizing compression, bottlenecking and growth control techniques in the architecture. Further, in Touvron et al. [2022] image patches are processed via linear sublayers operating across patches and channels, equipped with residual connections and forming image-level predictions with average pooling.

After the success of transformers in NLP tasks, they have been utilized in the visual domain with similar success. In Dosovitskiy et al. [2021] a "**Vision Transformer**" (**ViT**) is used directly on raw patch sequences as well as CNN embeddings, using a classification head for Imagenet data categorization. The work in Liu et al. [2021] hierarchically merges image patches (like approaches in Lazebnik et al. [2006]) into larger visual tokens, computing self-attention to shifting windows rather than the entire token set, which achieves linear complexity and comparable performance. A similar approach is pursued in Chen et al. [2021], where image patches at different scales are encoded separately, and subsequently fused with cross-attention strategies of different granularity between encoded tokens of each scale.

Regarding deep approaches for audio, an established technique has been to first convert audio content into images and then apply a visual DRL pipeline. This conversion usually involves spectrograms, i.e., 2D time and frequency representations, produced by applying short-time Fourier transform [Sejdić et al. 2009] on segmented audio clips. Applying CNN-based architectures has been a broadly popular approach, in a manner similar to vision DRL models. For example, in Hershey et al. [2017], the authors perform a large-scale evaluation of popular CNN architectures designed for the visual domain [He et al. 2016; Krizhevsky et al. 2012; Simonyan and Zisserman 2015; Szegedy et al. 2015] on audio spectrograms, introducing variation to the amount of data and labelset size for each experiment. All visual models considerably outperform a dense NN baseline applied on raw spectrogram features on audio classification. in Nanni et al. [2021], ensembles of popular CNN models with different data augmentations (signal, spectral, visual, and so on.) are applied on spectrograms and evaluated at multiple benchmarks. Furthermore, the authors of Wang and Oord [2021] adopt a contrastive learning approach, using waveform and spectrogram-level representations in a siamese configuration. They employ CNNs and Residual architectures on MFCC/spectrogram and raw audio inputs for representation learning, applying fixed features on sound event classification tasks.

Approaches like deconvolution [Zeiler and Fergus 2014] reverses CNN computation to project filterbank activations back to the pixel space. This is exploited in Choi et al. [2016] towards reproducing audio corresponding to weights of trained CNN networks, aiming at improving explainability of learned features. They utilize an architecture of convolution, pooling, and fully-connected layers on audio spectrograms for music genre classification. Transformer models are also utilized, e.g., in Gong et al. [2021], where a transformer encoder pretrained on Imagenet is used on input audio on spectrograms, classifying output *[CLS]* embeddings to auditory classes and outperforming convolutional architectures.

3 STRUCTURED KNOWLEDGE

Having covered representation approaches in the literature, we now move to examine structured knowledge resources available in the research community, that can be utilized to enrich ML systems in the context of classification tasks.

3.1 The Need for Enrichment

A classification machine has to overcome multiple difficulties in order to facilitate efficient and robust classification. One of the key challenges is clarifying ambiguity and deducing missing information in the input, the lack of which may hinder its decision-making abilities. Additional obstacles, limitations, and challenges that can arise may include:

- Critical contextual knowledge, disambiguating factors and crucial information that may be missing from training data: e.g., language ambiguity in text, unclear scale/orientation and pareidolia in audiovisual media
- Incomplete domain-specific knowledge in the training data, such as named-entity information, audiovisual logos and/or artifacts of special importance and/or contextual meaning.
- Factors inherent in the data generation setting. For example, subjective human attributes like education, writing style, cultural elements (prosody, dialects), shifting zeitgeist expressed in language/media (memes, slang, and so on.).
- Inherent data ambiguity (e.g., language polysemy, optical/auditory illusions, leading to diverse interpretations)
- Explainability/transparency standards in classifier outputs or internal operation: these may be required e.g., in use cases with both experts (e.g., in research, medicine, governance) and laymen (e.g., commercial products)
- Technological limitations, spanning a large semantic gap between engineered and real concepts [Sikos 2017]

In this survey, we explore promising methods that inject structured, curated knowledge to classifiers [Silvestri et al. 2021], mined from knowledge repositories, databases, ontologies and lexicons. Early research efforts relied on rule-based systems [Anaya 2011] that maintained expert knowledge preconfigured into decision rules relevant to the task, serving as a decision-making oracle. or built knowledge databases, inferring relationships and conceptual information by human experts [Cadoli and Donini 1997; Deng 1990]. This effort, along with the drive to digitize human knowledge led to a wealth of information available in machine-readable format. As a result, a large number of resources can be readily exploited towards improving tasks for which a relevant and applicable resource exists [Aye et al. 2008].

3.2 Knowledge Resources

We move on to discuss an indicative but diverse set of resources usable for enriching representations for classification. Table 2 provides details such as information unit types, relational structure, compilation means, and other information.

A popular resource is Wordnet [Miller 1995], a manually compiled **directed acyclic graph** (**DAG**) where nodes contain sets of synonymous semantic concepts ("synsets") along with example lexicalizations. Nodes connect with relations such as hypernymy/hyponymy (e.g., *dog* "is-a" animal) and meronymy (e.g., *wheel* "is-part-of" *car*) and additional metadata (e.g., examples and definitions). Sentiwordnet [Baccianella et al. 2010] is a related resource having automatic synset annotations with positive/negative/neutral polarities, scored in $[0.0, 1.0]$ and adding to unity. Babelnet [Navigli and Ponzetto 2010] is a network disambiguating and linking Wikipedia lemmas, encyclopedic content and synsets, while machine translation and cross-lingual links from Wikipedia provide multilingual information. Framenet [Baker et al. 1998] is a hierarchical lexical database annotated with descriptions of semantic roles pertaining to, e.g., events, relations, situations, and related entities, based on frame semantics [Fillmore 2006].

Further, CyC [Lenat et al. 1990] consists of a knowledge base that catalogues numerous formal assertions that represent existing human knowledge. Facts are encoded into a handcrafted

Table 2. Knowledge Resources for Enrichment of Classification Tasks

name	unit type	relation	compilation	endpoint / format	url	language
Wordnet [Miller 1995]	concept	hierarchical, semantic	manual	multiple / multiple	web , nltk	multiple
SentiWordnet [Baccianella et al. 2010]	concept	polarity, hierarchical, semantic	automatic	python, web / text	nltk, code	English
Framenet [Fillmore et al. 2003]	semantic frame	frame-semantic	manual	python / -	web	multiple
Babelnet [Navigli and Ponzetto 2010]	concept / entity	hierarchical, semantic, similarity	automatic	multiple / multiple	web	multilingual
DBpedia [Lehmann et al. 2015]	property-value	linked-data	mixed	web, SPARQL, REST / RDF, JSON	web, code	multiple
Wikidata [Vrandečić and Krötzsch 2014]	property-value	linked-data	manual	multiple / multiple	web	English
ParaphraseDB [Ganitkevitch et al. 2013]	phrase	paraphrasal	automatic	web / text	web	multiple
Freebase [Bollacker et al. 2008]	property-value	linked-data	manual	web / text	web	English
Probase [Wu et al. 2012]	concept	hierarchical, semantic	automatic	-	web	English
ASER [Zhang et al. 2020]	phrase	semantic, causal	automatic	REST, python	web, github	English
YAGO [Suchanek et al. 2007]	entity	hierarchical, semantic	manual	REST, web / JSON	web	multiple
ConceptNet [Liu and Singh 2004]	concept	hierarchical, semantic, similarity	mixed	python, REST, web / JSON	web	multiple
CyC [Lenat et al. 1990]	concept / entity	hierarchical, semantic	manual	multiple / multiple	web, code	English
Imagenet [Deng et al. 2009]	concept	hierarchical	manual	web / XML	web	English*
Visual Genome[Krishna et al. 2017]	concept	semantic	manual	multiple	web	English*
Audioset [Gemmeke et al. 2017]	concept	hierarchical	manual	web / JSON	web, web	English*
Music Ontology [Raimond et al. 2007]	concept	semantic	manual	multiple / multiple	web	English*
Audio Feature Ontology [Allik et al. 2016]	concept	semantic, hierarchical	manual	RDF	web	English*
COMUS [Rho et al. 2009]	concept	semantic, hierarchical	manual	RDF	web	English*
E-ANEW [Warriner et al. 2013]	lexicon	affective	manual	web / csv	web	English
General Inquirer [Russell 1980]	lexicon	affective	manual	java, python / -	java, python	English
Labelsets	concept	hierarchical, semantic	-	-		-

For programmatic resource endpoints and labelsets, the format is irrelevant and is omitted. In the language column, entries with a asterisk superscript* refer only to the language in which the resource elements are described in – i.e., the resource content itself is not linguistic and does not lend itself to a specific language.

knowledge representation language (CycL) using a logical framework, forming an ontology and inference engine developed by experts. The ConceptNet [Liu and Singh 2004] graph database captures knowledge via semi-structured natural language fragments, using nodes that represent both simple concepts and compound entities and events, built by combining primitive building blocks like verbs, noun and prepositional phrases (e.g., *eat lunch*, *in the evening*). Edges map relations like synonymy, meronymy, causality and affect, as well as probabilistic associations (e.g., “it often holds that”), corresponding to “informal everyday knowledge”. It was built by rule-based extraction on crowdsourced “fill-in-the-blank” commonsense tasks. Probase [Wu et al. 2012] is a large hierarchical taxonomy of 2.7 million concepts, automatically built by parsing large volumes of webpage content. It includes a probabilistic model to encapsulate ambiguity, degrees of certainty and learned inconsistencies from web source content. ASER (activities, states, events and relations) [Zhang et al. 2020] is an eventuality knowledge graph parsed from text data. Eventualities are verb phrases, linked with temporal, contingency, comparison, expansion and co-occurence relations, with the resulting graph conveying rich semantics.

DBpedia [Lehmann et al. 2015] utilizes Wikipedia article infoboxes, manual, and automatic rule-based extraction, named-entity recognition, and statistical methods to build a semantic web structure, using linked data technologies that facilitate cross-lingual cataloguing of Wikipedia literals and property-based relations. Wikidata [Vrandečić and Krötzsch 2014] is an open collaborative platform aiming at providing the data available on Wikipedia in a structured and easy-to-use format. It provides property-value pairs (e.g., *author - George Orwell*) as well as more complex contextual relationships via the use of “qualifiers” (e.g., conveying temporal information). The YAGO project [Suchanek et al. 2007] combines Wordnet and Wikipedia information via automatic heuristics, finalized by quality control-oriented post-processing. An ontology links entities with each other (with semantic, linguistic, hierarchical and real-world associations like “hasWonAward”) and with literals (e.g., to string lexicalizations and numbers). Freebase [Bollacker et al. 2008] is a large collaborative database consisting of entities, properties and assertions, organized as tuples in a graph data store. It emphasizes on scalability, collaborative maintenance and ease of use, towards research and open data-oriented community applications. Furthermore, ParaphraseDB [Ganitkevitch et al. 2013] is a database containing paraphrase pairs, i.e., pairs of semantically equivalent, but syntactically and/or lexically different phrases. Pairs are built by analyzing parallel corpora with bilingual pivoting: paraphrases are source pairs that translate to the same string in a foreign language. Subsequent refinements with distributional re-ranking take into account contextual information.

Multiple resources are organized in lexicon/name-value formats. Expanding on previous work [Bradley and Lang 1999], E-ANEW [Warriner et al. 2013] contains affective norms like *valence*, *arousal* and *dominance*, referring to degrees of pleasantness, emotion and power/control, over 14 thousand English words. Additionally, the General inquirer [Russell 1980] psycholinguistic model maps affect in spatial coordinates for over 8 thousand English words, mapped to 182 affective categories that emphasize and focus on coverage of a broad range of psychological states.

There are some knowledge resources that deal with multimedia. Imagenet [Deng et al. 2009] links 80 thousand Wordnet synsets to hundreds of high-resolution representative images, including animals, objects, scenes, and so on, and bounding box annotations, for a total of 3.5 million images. The data was compiled through automated retrieval methods and refined by quality control via crowdsourcing. Visual Genome[Krishna et al. 2017] provides 100 thousand images with dense crowdsourced annotations such as object regions, attributes, relationships, and scene graphs. Additionally, region descriptions, question-answer pairs and Wordnet linkage are included, towards "grounding visual concepts to language". Further, the Audioset ontology [Gemmeke et al. 2017] organizes audio categories into a hierarchical structure with abstract classes such as "human sounds", "music", "natural sounds" and refined categories like "whistling", "musical instrument" and "wind". Concepts come with a short textual description and representative multimedia links. The Music Ontology [Raimond et al. 2007] involves musical concepts in three levels of granularity and expressiveness, from editorial (e.g., track/artist/album information), through performance-related concepts (e.g., performance, recording, and audio stream-level attributes), to decomposition to fine-grain musical elements (e.g., key, musical instrument, and temporal localization of an audio piece). An extension of the above work is the **context-based music recommendation** (**COMUS**) ontology [Rho et al. 2009]. It consists of hierarchical, music-related relationships, definitions and attributes, associating them with people, genre, mood states, locations, situations, and events. The Audio Feature Ontology [Allik et al. 2016] provides multiple levels of abstraction for describing the audio feature extraction process, containing a catalogue of ≈400 features. Entries are organized with respect to data density and temporal characteristics, ranging from abstract conceptualizations (e.g., "chromagram") to specific extraction algorithms.

Alternatively, exploitable knowledge can be mined from the ground truth/labelset of the data, in a modality-agnostic way. For instance, a low confidence for a superclass in a label hierarchy automatically provides a bias against predicting its children. This can be exploited by, e.g., hierarchical classification [Silla and Freitas 2011], but also serves as knowledge that can be integrated in representations of the data in question. In addition, ground truth/metadata can be utilized in the representation construction phase, such as user tags and textual descriptions of multimedia instances.

Having presented a set of knowledge resources capable of enhancing classification, we move on to present related work that realize this, implementing various enrichment approaches for different data modalities.

4 REPRESENTATION ENRICHMENT APPROACHES FOR CLASSIFICATION

Here we combine the material covered into previous sections – i.e., content-based approaches from Section 2 and knowledge resources/representation enrichment motivation from Section 3— to a presentation of methods in the literature that inject knowledge into data representations, in order to enhance classification tasks for different modalities.

We separate and organize the outlined work in three broad categories, depending on how the enrichment is applied in the representation construction process. At the same time, we draw parallels to the content-based categories outlined in Section 2, to which this grouping bears meaningful analogies. Given this context, the enrichment paradigms include:

- **Input enrichment/modification (IEM)**, covering approaches that insert external knowledge at the input feature level, arriving at a configuration where knowledge-based information is fed as an input of the classification pipeline. The simplicity of this paradigm bears similarities to **low-level/template-matching approaches** (**LLTM**), since knowledge is treated as an auxiliary modality/channel to data content.
- **Knowledge-based refinement (KBR)**, which includes methods that transform/modify/process existing low-level representations in a preconfigured manner that is directed, guided and/or quantified by external information. This approach builds upon the **aggregation-based** (**AGB**) paradigm for content-based representations, with the critical component of knowledge influencing the aggregation mechanism.
- **Knowledge-aware end-to-end (KAE)** systems, where deep hierarchical architectures are built from both content-based and knowledge-based information. The paradigm expands upon **deep representation learning** (**DRL**) content-based models, by also including knowledge for building enriched feature hierarchies.

The following sections elaborate on each enrichment category: Section 4.1 deals with IEM methods, followed by KBR and KAE systems, in Sections 4.2 and 4.3, respectively. Covered studies are presented in Table 3, conveying similar information to the corresponding table in Section 2; here we additionally include the enrichment strategy (IEM/KBR/KAE) and the information about the knowledge resource and/or type. While non-exhaustive, the presented body of work constitutes a descriptive set of characteristic approaches for each proposed enrichment paradigm.

4.1 Input Enrichment and Modification

4.1.1 Overview. A straight-forward avenue towards knowledge-rich approaches for classification is injecting knowledge alongside the content-based feature set C extracted from the input data. This involves modifying the collection of features fed to the learning model with high-level semantic, statistical, and conceptual information S. Such information is obtained by mining structured knowledge resources to extract knowledge relevant to the input instance, resulting in an enriched feature set $E = f(C, S)$. The modification function f performs the fusion, e.g., by applying operations such as rule-based selection, concatenation, replacement, and so on.

4.1.2 Approaches. A major component in IEM is selecting the resource that will provide the knowledge-based features, as these will merged to content-based features and be directly used by the downstream classifier.

Regarding text, resources such as Wordnet have been widely used for directly extracting conceptual/sense-level information from lexicalizations as well as exploiting hierarchical structures in its graph. Such schemes yield common features for different texts with similar semantics, improving generalization potential for the downstream learning machine. For example, the authors in Elberrichi et al. [2008] combine BoW with Wordnet sense statistics weighted via TFIDF. They explore different ways for mapping lexicalizations to senses (e.g., including expanding matches to hypernyms), concatenating the lexical and semantic channels and applying χ^2-based feature selection. A similar approach in Nezreg et al. [2014] improves upon Wordnet concept lookups by utilizing multi-word querying and POS information from Treetagger [Schmid 1994]. Further, conceptual mappings are expanded by considering hierarchical relations (e.g., hypernymy,

Table 3. Indicative Studies using Representation Enrichment from External Information Sources for Classification

citation	mod.	enrichment - resource	category	representation	labelling	classifiers	metrics
[Elberrichi et al. 2008]	TXT	IEM - Wordnet	LLTM	TFIDF	MC-SL/ML	similarity	F1
[Kumar and Minz 2013]	TXT	IEM - SentiWordNet	AGB	TFIDF, PCA, LSA, GI, GR	MC-SL	SVM, NB, k-NN	ACC
[Nezreg et al. 2014]	TXT	IEM - Wordnet	LLTM	TFIDF	MC-SL/ML	SVM, DT, kNN	P
[Pittaras et al. 2020]	TXT	IEM - Wordnet	DRL	CBOW, BoW, TFIDF	MC-SL/ML	MLP	F1
[Škrlj et al. 2020]	TXT	IEM - Wordnet	LLTM	BoW, TFIDF	MC-SL	SVM, MLP, LSTM	F1
[Škrlj et al. 2021]	TXT	IEM - ConceptNet	LLTM	TFIDF, Word2Vec	MC-SL	SVM, NEURAL, LR	ACC
[Li et al. 2017]	TXT	KBR - Sentiment lexicon, synonyms	LLTM	GloVe	MC-SL	SVM, Adaboost, NB	F1
[Yu et al. 2017]	TXT	KBR - E-ANEW	DRL	skipgram, GloVe	MC-SL/ML	CNN, DAN, LSTM	ACC
[Glavaš and Vulić 2018]	TXT	KBR - Wordnet, synonyms/antonyms	DRL	skipgram, GloVe, fasttext	MC-SL	similarity	ACC
[Shi et al. 2019]	TXT	KBR - Paraphrases	DRL	ELMO	BIN	MLP	ACC
[Chen et al. 2018]	TXT	KAE - Wordnet	DRL	BILSTM, GloVe	MC-SL	NEURAL	ACC
[Sun et al. 2019b]	TXT	KAE - Entities (Dataset)	DRL	TRANSFORMER	BIN	NEURAL	ACC
[Zhang et al. 2019b]	TXT	KAE - Entities (TAGME, Wikidata)	DRL	TRANSFORMER	MC-SL	NEURAL	ACC
[Peters et al. 2019]	TXT	KAE - Entities (Wikipedia, YAGO, Wordnet)	DRL	TRANSFORMER	BIN/MC-SL	NEURAL	ACC
[Ke et al. 2020]	TXT	KAE - Sentiwordnet	DRL	TRANSFORMER	MC-SL	NEURAL	ACC
[Li et al. 2022]	TXT	KAE - Metadata	DRL	TRANSFORMER	MC-SL	NEURAL	ACC
[Liu et al. 2022]	TXT	KAE - Probase	DRL	CNN, TRANSFORMER	MC-SL	NEURAL	ACC
[Benitez and Chang 2003]	IMG	IEM - Wordnet, tags	AGB	TFIDF, color, KMeans, k-NN, SoM	MC-SL	NB, SVM, BN	ACC
[Vogel and Schiele 2007]	IMG	IEM - Labelset	LLTM	color, direction, signal	MC-SL	SVM	ACC
[Marszalek and Schmid 2007]	IMG	IEM - Wordnet, tags	AGB	SIFT, KMeans	MC-SL	SVM	ROC
[Kliegr et al. 2008]	IMG	IEM - Wordnet, tags, Wikipedia	LLTM	MPEG-7	MC-SL	evol. SVM	P, R
[Binder et al. 2009]	IMG	IEM - Labelset, VOC2006	AGB	SIFT, KMeans, SPM	MC-SL	SVM	ACC
[Li and Sun 2006]	IMG	KBR - Wordnet	AGB	color, texture, shape, KMeans	BIN,MC-SL	LM, SVM	P, R
[Wu et al. 2010]	IMG	KBR - region	AGB	SIFT, SPC	MC-SL	SVM	AUC
[Deselaers and Ferrari 2011]	IMG	KBR - Imagenet	LLTM	GIST	MC-SL	similarity, SVM	ROC
[Li et al. 2014]	IMG	KBR - Metadata	DRL	Decaf, TF	MC-SL	SVM	mAP
[Menglong et al. 2019]	IMG	KBR - Labelset	DRL	CNN	MC-SL	NEURAL	ACC
[He et al. 2021]	IMG	KBR - Labelset	DRL	CNN	MC-SL/ML	NEURAL	ACC
[Marino et al. 2017]	IMG	KAE - Wordnet, Visual Genome	DRL	CNN	MC-SL	NEURAL	mAP
[Zhang et al. 2019a]	IMG	KAE - Wordnet, Imagenet, Labelset	DRL	CNN	MC-SL	NEURAL	ACC
[Li et al. 2019]	IMG	KAE - Labelset	DRL	TRANSFORMER, CNN, DNN	MC-SL	NEURAL	ACC
[Noh et al. 2019]	IMG	KAE - Wordnet, Visual Genome	DRL	TRANSFORMER	MC-SL	NEURAL	VQA
[Jayathilaka et al. 2021]	IMG	KAE - Labelset	DRL	CNN, MLP	MC-SL	NEURAL	ACC
[Yang et al. 2022]	IMG	KAE - Wordnet	DRL	GNN, PCA, KMeans	MC-SL	NEURAL	ACC
[Cano et al. 2004]	AU	IEM - Labelset	LLTM	MFCC, spectral, psych.	MC-SL	k-NN	ACC
[Hu and Downie 2010]	AU	IEM - Wordnet, E-ANEW, GI	LLTM	TFIDF, signal, spectral, MFCC	MC-SL	SVM	ACC
[Jamdar et al. 2015]	AU	IEM - E-ANEW, Wordnet	LLTM	psych., musical	MC-SL	k-NN	ACC
[Cheng et al. 2008]	AU	KBR - Labelset	AGB	PCP, LCSS, MFCC, ngrams	MC-SL	k-NN	ACC
[Pachet and Roy 2009]	AU	KBR - signal proc.	LLTM	signal, spectral	MC-SL	kSVM	ACC
[Favory et al. 2020]	AU	KBR - Metadata	DRL	CNN, DNN, autoencoders	MC-SL	MLP	ACC
[Zharmagambetov et al. 2022]	AU	KBR - Labelset	DRL	CNN, LSTM	MC-SL/ML	NEURAL	F1
[Bertero and Fung 2016]	AU	KAE - Wordnet, SentiWordnet, Metadata	LLTM, DRL	MFCC, spectral, signal, psych., word2vec	MC-SL	k-NN	ACC
[Jiménez et al. 2018]	AU	KAE - Labelset	DRL	CNN, SIAMESE	MC-SL/ML	NEURAL	ACC
[Sun and Ghaffarzadegan 2020]	AU	KAE - Labelset	DRL	CNN, LSTM	MC-SL/ML	NEURAL	F1
[Zhang et al. 2021]	AU	KAE - ASER	DRL	GCN	MC-SL/ML	NEURAL	mAP

Notation follows Table 1 – additionally, the enrichment-resource column contains entries in the form ENR-RES, where ENR refers to enrichment approaches described in Section 4 and RES to the resource(s) utilized for knowledge injection.

antonymy) and different concept weighting schemes are investigated for the final concatenation of enriched TFIDF features. Further, the work in Pittaras et al. [2020] combines Wordnet sense statistics and CBOW embeddings [Mikolov et al. 2013] over different lexico-semantic fusion, disambiguation and concept weighting methods. Concept mapping is expanded and regularized with hypernymy-based spreading activation [Collins and Loftus 1975], diffusing semantics in a controlled manner per expansion step. Moreover, document-level taxonomies are built by utilizing hypernymy in Škrlj et al. [2020], leveraging different lexical BoF/TFIDF vector configurations for different terms. Semantic vectors are built via double-normalized TFIDF [Manning et al. 2008] after sense disambiguation, followed by statistical, graph-theoretic and ranking-based feature selection to arrive at a fixed dimensionality.

Additional resources exploited include SentiWordNet for sentiment mining in Kumar and Minz [2013]. The authors extract sentiment scores from mined senses, using TFIDF and various content-based aggregation/selection means. The work in Škrlj et al. [2021] mines ConceptNet to generate semantic relational features, by grounding triplets in the ontology during document traversal and collecting TF-IDF bags of relations for each document. These are subsequently concatenated with different token-level and distributed features for multiple classification tasks.

For images, many works employ relevant ground truth like class hierarchies, label semantics, and localized annotations to create high-level representations with IEM, given a lack of high-level visually-oriented knowledge resources.

For instance, visual annotations to image regions and segments have been exploited as prior semantic knowledge. Such metadata can be utilized for high-level semantic feature generation in classification tasks in a variety of ways, and are usually available as supplementary ground truth information in the Labelset. Early applications [Luo and Savakis 2001] use regions marked as "sky" or "grass" to build intermediate classifiers as semantic feature generators, combined with color and texture features. An array of pre-selected concept annotations to fixed-size regions is used in Vogel and Schiele [2007], fed to classifiers that generate conceptual model vectors as high-level features that are concatenated with multiple low-level features such as color, direction and intensity histograms.

Another useful source of knowledge in image datasets are tags, conveying high-level semantics exploitable for improving performance. Utilization avenues include adopting representation extraction techniques for text, using responses as higher-level semantic features. Given this textual modality data, additional conceptual information can be obtained by probing NLP-oriented knowledge resources such as Wordnet. To this end, the work in [Benitez and Chang 2003] use aggregated color features along with image tags mapped to TF and TFIDF vectors. Multimodal vectors are clustered into knowledge graph of mid-level concepts with various techniques, used as high-level semantics. Wordnet is used for sense disambiguation and facilitation of semantic similarity extraction between graph concepts. Additionally, the authors in Marszalek and Schmid [2007] use textual tags, annotations and labels to extract training examples from Wordnet, exploiting its relational structure (e.g., hypernymy, meronymy, and holonymy) with pruning applied to inhibit propagation towards too generic graph nodes. Sense activations are then combined with SIFT features and a visual BoW approach. Segment predictions are enhanced with tag-based semantics in Kliegr et al. [2008]: first, a **named entity recognition** (**NER**) procedure probes Wordnet to discover concepts most similar to an image tagset, via a targeted hypernym discovery process and special measures to improve coverage (e.g., by retrieving relevant Wikipedia articles in terms of textual similarity and article linkage). This procedure results in a collection of senses matching image tags, which are fused with MPEG7 features on regions segmented via self-organizing maps [Kohonen 1990] and particle swarm optimization [Kennedy and Eberhart 1995]. Other approaches include building taxonomy membership features, expressing instances with respect to class memberships in a taxonomic resource. For example, the work in Binder et al. [2009] utilizes the VOC2006 taxonomy [Everingham et al. 2005], exploiting hypernymy-based node membership of the image class in the taxonomy graph to build taxonomic features and affect the training loss by modifying the prediction target, in conjunction with visual features based on SIFT and visual BoW.

Regarding audio data, a similar setting persists, as the visual domain: namely, auxiliary metadata such as ground truth information and textual annotations are the facilitator of enrichment with an IEM strategy.

One such avenue can be found in the music classification domain, where text metadata in the form of song lyrics are exploited for enrichment, via taking advantage of IEM methods for text. Such an approach is adopted by Hu and Downie [2010], utilizing audio and text features: Wordnet, Wordnet-Affect [Strapparava et al. 2004], General Inquirer and E-ANEW provide semantic and psycholinguistic responses, generating statistical, POS and heuristic-based sense and text features in TF and TFIDF ngram bags. These are used to enrich of MFCC, spectral and statistical audio features in various combination configurations. Additionally, in Jamdar et al. [2015] E-ANEW is used to mine valence and arousal scores from song lyrics, with Wordnet powering mapping expansion and semantic disambiguation via sense synonyms and POS information, respectively. Knowledge-based features are adjusted to song-level values and combined with multiple musical, rhythmic and psychoacoustic features, followed by normalization and scaling.

Other approaches exploit relations found in the ground truth and/or labelset of the available audio data. Namely, an early approach uses coarse-grained classification as intermediate high-level features in Zhang and Kuo [1998], with audio-oriented concept scores being used as model vectors, subsequently classified into finer target classes (e.g., rain, bird sounds). Audio content is modeled via signal and spectrum statistics, psychoacoustic/musical features and rule-based clustering. The study in Cano et al. [2004] uses audio samples with Wordnet annotations expanded with additional audio-related concepts. This ground truth is utilized for training a classification system to produce concept-level confidence scores, such as producing assignments to musical instruments. MFCC, statistical, spectral and psychoacoustic features are utilized for the representation, fed to k-NN classifiers for instrument sound categorization.

4.2 Knowledge-based Refinement

4.2.1 Overview. Here we present a knowledge-guided analog to AGB, modifying features in an informed manner. In KBR, knowledge-based relations can be used as aggregation criteria for similar operations on corresponding content-based features, acting as, e.g., filtering and membership indicators for informed data fusion. Moreover, representation learning methods can inject constraints, regularization, and scaling in the loss/similarity function or fitness objective. For example, taxonomic information linking the concepts "dog" and "wolf" (i.e., subspecies of "Canis") may be exploited by an embedding generator to bring such data points closer in the embedding space. Likewise, a database of synonyms could make a BoW approach to merge weights matched to concepts "dog" and its synonym "Canis Familiaris". Such modifications are informed/quantified by utilizing external knowledge of suitable encompassed relationships.

4.2.2 Approaches. KBR methods for text exploit relations in NLP-related resources to influence representations. One specific approach utilizes word groups, based on categorical assignments, numerical scores or pairwise relations. Intra-group member vectors are then biased to lie close in the embedding space during training or post-processing phases.

A number of works follow this procedure with sentiment-oriented knowledge. For instance, the approach in Li et al. [2017] expands the GloVe model of Pennington et al. [2014] to apply sentiment-guided bias on the word-level, via positive/negative/neutral tags obtained by lexicons with synonym expansion, and globally, via document-level sentiment annotations. Moreover, the work in Yu et al. [2017] introduces refinement of word vectors in a two-stage re-ranking scheme: first, the 10 nearest neighbors of a word are retrieved, followed by re-ranking with respect to the degree of positive and negative sentiment expressed by each. The cosine similarity between word vectors is used to obtain semantic similarity, while the sentiment score is provided by the "valence" norm in the E-ANEW [Warriner et al. 2013] lexicon. The refined output of GloVe and Skipgram vectors then used as the final representation.

Other semantic relationships exploited are paraphrase pairs, i.e., text tuples of different contents but synonymous semantics. Namely, the work in Shi et al. [2019] retrofit contextualized ELMO embeddings [Peters et al. 2018] with paraphrasal information to learn an orthogonal transformation that maps paraphrasal pairs to collocated projections targets. Further, semantic word relationsips of synonymy and antonymy are utilized in works such as in Glavaš and Vulić [2018]. There, the authors seek to optimize with respect to antonymy and synonymy constraints while trying to maintain distances between remaining instance pairs. The investigation explores different optimization objectives, knowledge resources (i.e., Wordnet and Roget's Thesaurus [Jarmasz and Szpakowicz 2004]) and source embeddings (Skipgram, GloVe, and FastText [Joulin et al. 2017]) to train a mapping with a dense neural network.

Regarding images, one KBR technique exploits metadata with text-oriented enrichment methods, where graph-based resources like Wordnet are useful since they provide multiple exploitable relations via node linkage. For example, early work in Srikanth et al. [2005] uses region descriptions, color, position, texture, shape, and blob-to-word translation associations. They build an "ontology-induced visual vocabulary" by reweighing KMeans with word contributions per cluster, while generative region modeling to Wordnet senses is used to facilitate classification. A similar approach in [Li and Sun 2006] builds a KMeans codebook from keyword-annotated regions, a process augmented with constraints from Wordnet relations that modify the clustering objective to consider the semantic similarity between keywords.

Additional knowledge resources have been exploited to modify representations in the visual domain, such as Imagenet, which provides a visual component to the hierarchy of the Wordnet semantic graph. It is exploited in Deselaers and Ferrari [2011] for computing semantic similarity by extracting the k visually nearest neighbors in Imagenet for each image in an input pair, using a similarity measure based on GIST [Oliva and Torralba 2001], with a final score computed either from pairwise neighbor similarities or the overlap of their category distributions.

Additionally, region-level ground truth and high-level semantic model vectors have been use leveraged in KBR methods. First, image region annotations specifying whole individual objects can instruct representation systems to pool together local descriptors extracted from that region. This is investigated in Wu et al. [2010], where object and foreground annotations bias features extracted from similar regions to be grouped together to same visual words, in a **"semantics-preserving" codebook** (**SPC**) process. SIFT with SPC and different distance metric learning techniques such as **neighbourhood component analysis** (**NCA**) [Goldberger et al. 2004] are used to build the final representation. Second, a model vector approach may leverage predictions of state-of-the-art classifiers to produce high-level semantics and structures usable for defining meaningful relationships for subsequent representation refinement. Such an approach is pursued by Menglong et al. [2019], using popular CNNs to build a knowledge graph and marking the top k classification results as related, forming pairwise category relationships. Content-based results are refined with a graph-based category similarity measure, improving the accuracy of different deep convolutional nets.

Another refinement avenue approach includes exploiting knowledge-enabled expansion of the available data pool, building the resulting representations with the support of additional data. For instance, in Li et al. [2014], the authors adopt a multi-instance learning approach with "privileged information" [Vapnik and Vashist 2009], i.e., additional extraneous information available during training. This is realized by using retrieval methods to obtain additional related training samples from the web. For training, metadata in the form of image text descriptions are exploited and to TF vectors, while convolutional Decaf features are used for visual content description [Donahue et al. 2014].

Regarding refinement approaches for audio data, existing methods utilize different subdomains of audio-related knowledge in a variety of forms, to produce enrichment via feature refining procedures. For instance, knowledge related to audio signal processing presents one such avenue: this approach is followed in Pachet and Roy [2009], where the "Analytical Features" framework encapsulate knowledge over audio feature engineering and design paradigms (e.g., different practices, heuristics and patterns). These are encoded into operators that process/transform/compose different signal/spectral-based information. The framework is evaluated over fine-grained categorization for artificial and natural sounds (e.g., dog barks, percussion) with a polynomial SVM.

As in other modalities, ground truth presents a rich source of knowledge in audio data. For instance, fine-grained musical annotations can be utilized as conceptual features and high-level representations, which is explored in Cheng et al. [2008] in the form of manually annotated chord transcriptions. These are used to train a chord-ngram HMM model for tasks such as music emotion classification, which captures sequential information of detected chord progressions, along content-based features such as PCP vectors, Longest Common Chord Subsequence similarity, histogram-based measures and MFCC. The work in Zharmagambetov et al. [2022] uses a CNN - LSTM architecture [Hochreiter and Schmidhuber 1997] jointly with a tree-based ontology built from the training labelset. The ontology is used in a decision tree structure to explicitly model the hierarchical probability distributions during learning, which additionally utilizes unlabeled data via consistency training. Further, in He et al. [2021], hierarchical ontologies are manually built by considering dataset and labelset-derived semantics, defining coarse and fine label separations. Ontology levels then determine batch sampling strategies for triplet generation, used to feed a CNN equipped with a triplet loss.

Other types of ground truth include higher level information, such as musical genre, which can be used as conceptual features for generic audio classification. This is investigated in Favory et al. [2020], where an autoencoder scheme is used for aligning content and knowledge in a dual channel architecture. The first channel involves a content-based convolutional autoencoder that ingests spectrogram representations of the audio signal and is trained with a reconstruction loss. The second channel supplies one-hot encodings of semantic audio tags to a feed-forward NN, fitted via cross-entropy. The two learned representations are biased to align via a contrastive loss producing semantically enriched features, fed to an MLP for sound event recognition, instrument and genre classification tasks.

4.3 Knowledge-aware End-to-end Systems

4.3.1 Overview. Adhering to the popularity and performance of large end-to-end learning models, KAE approaches utilize knowledge in conjunction with deep learning and neural networks. Here, the injection of external information may occur as input and/or refinement operations, as in previous enrichment strategies—the distinguishing factor is that here, content-based and external information is jointly exploited in an end-to-end fashion, to automatically learn knowledge-enriched feature hierarchies. As in DRL, KAE approaches often jointly learn the representation and discrimination components that facilitate classification, and routinely arrive at highly transferable representations.

4.3.2 Approaches. A rich collection of knowledge resources have been utilized for KAE methods, leveraging unsupervised training on enriched data inputs for high-performance transfer learning. Entity information is one such knowledge domain, assigning high-level semantics to sequences of words and linking together different entity lexicalizations.

Entity-aware language models for text are investigated in knowBERT [Peters et al. 2019], providing approaches for integrating pretrained models with generic knowledge bases structured as triplets or graphs. Self-attentive mention-span embeddings are computed for each candidate entity, followed by neural entity linking [Kolitsas et al. 2018; Lee et al. 2017]. Lexical embeddings are then re-contextualized with the entity-span vectors, using a multihead attention transformer layer. Wikipedia, YAGO and Wordnet are examined for entity identification, with embeddings built via skipgram vectors. Another approach integrating entities along with phrase semantics is ERNIE [Sun et al. 2019b]. ERNIE expands the BERT model [Devlin et al. 2019] to consider knowledge base information, introduced by knowledge-level masking: phrase-level and conceptual/named-entity-level masking is adopted, extending the **byte-pair encoding** (**BPE**) used in BERT. This modeling approach is utilized in pretraining, which is performed on multiple-domains and heterogeneous data. Another model focusing on named entities [Zhang et al. 2019b] uses distinct textual and knowledge encoders to handle lexical/syntactic and fine-grained entity-related information, respectively: in contrast with previous studies, content and knowledge data is fed to the model via different input channels. TransE [Bordes et al. 2013] and Wikidata are used to encode entities into embedding vector, which are combined with lexical token embeddings via multi-headed attention, while TAGME [Ferragina and Scaiella 2010] is used for entity extraction. Entity prediction is added as a pretraining task, utilizing special tokens for entities and masked language modeling.

Furthermore, sentiment and semantic word relationships have been used in KAE; such information can highlight relations between word/sequence pairs, which can be exploited and learned in neural architectures. For instance, in SentiLARE [Ke et al. 2020], POS information is used to extract word-level sentiment polarity from SentiWordnet [Baccianella et al. 2010] to train a multilayer transformer. Masked language modeling is used for pretraining, using context-aware sentiment attention to weigh polarities from individual words to the sentence-level sentiment score. Pretraining subtasks include predicting both sentence and word-level labels. in Chen et al. [2018], lexical relations including synonymy, antonymy and hypernymy are mined from Wordnet in order to enhance premise / hypothesis classification on sentence pairs. The relations are represented as binary vectors, mapped via graph embedding methods (i.e., TransE). The model uses two biLSTM layers for sequence encoding and decoding word relationships, respectively. Semantic lexical information is weighted by the alignment score between word pairs. Moreover, the work in Liu et al. [2022] extracts conceptual information of input words from the Probase taxonomy [Wu et al. 2012], applied in a short text classification setting. Temporal CNN and transformer architectures are used for embedding and representation construction respectively, with the content-based and conceptual channels being merged via cross-attention to enriched features.

Finally, knowledge in the form of dataset metadata is exploited in Li et al. [2022] with the DASK system; it identifies domain-independent words from dataset source information and builds knowledge graphs that encapsulate their relationship to domain-related content. They use a BERT variant on knowledge-injected data to classify user reviews.

Ground truth and labelset information is widely used in KAE methods for images. For instance, the work in Yang et al. [2022] utilizes content-based information with neighborhood embeddings and knowledge from Wordnet via vectorizing textual node descriptions, building a unified knowledge graph. The structure is sampled by graph attention modules for few-shot classification, using separate subgraphs for different tasks for debiasing. Visual Genome has been exploited to construct fitting knowledge structures for task-specific enrichment. A popular such format is visual knowledge graphs, utilized for enriched image classification. For instance, Visual Genome has been used to build a knowledge graph of candidate image labels in Marino et al. [2017], using encapsulated object-object/object-attribute relationships and fusing scene graphs with Wordnet semantics. Classification uses a Graph Search Neural Network [Li et al. 2016], that scores each node in the knowledge graph and provides a global aggregate prediction. An object detector/classifier is used to identify an initial visual node in the image, from which controlled propagation supplies additional neighbor nodes of visual objects. A visual knowledge graph is also built in Zhang et al. [2019a], containing semantic associations between content in the image to objects and scenes relevant to it. Scene labels are produced either by classifiers trained on Imagenet, or semantic associations of the image lexical label, which are extracted from Wordnet. A similarity score based on co-occurrence statistics of objects/scenes detected between images pairs is used to augment learning, fed to different CNNs to predict object/scene labels. Ground truth in the format of scene graphs has also been used in [Li et al. 2019] for **visual question answering (VQA)**, which can be viewed as answer classification over visual and textual inputs. Neural perceptual modules (convolutional, attention-based and dense) are used as specialized operators, each adhering to specific semantic subtasks (e.g., boolean operations, localization of salient/relevant regions, inference) matching the question type structure. Scene graph annotations consisting of object region coordinates, attributes and relations are used to guide the layout generation and optimization, propagating through each module.

Further, ground truth from Visual Genome is exploited in Noh et al. [2019] for VQA tasks. First, structured knowledge in the form of visual descriptions and answers/labels from Visual Genome is used to generate blanked image descriptions that characterize visual recognition tasks. To disambiguate candidate tasks, Wordnet sense-based modeling is used: sampling a task specification during training involves retrieving senses with large lexicalization overlaps with the input question. Sampled task-conditional visual classifiers are subsequently used to score candidate answers, using pretrained neural models and attention-based modeling. In Jayathilaka et al. [2021], explicit and inferred pairwise hierarchical label relations (e.g., subsumption, disjointness) are mapped to n-ball conceptual embeddings. They are joined with content-based DCNN features and projected to the conceptual space with an MLP, with the learning process mapping visual features into the conceptual space defined by the ontology embeddings.

Regarding enrichment of audio classification with KAE systems, a variety of knowledge exploitation methods have been tried. As in the visual modality and other enrichment avenues, text-based ground truth has been a popular choice and important knowledge contributor. For instance, in Bertero and Fung [2016], audio data and text transcriptions are used in a multimodal approach, employing MFCC, spectral, signal-based and psychoacoustic features for audio representation. For text, content is mapped to Word2Vec embeddings, bags of ngrams and features related to syntax, sentiment, antonyms, and speaker turn. Wordnet and SentiWordnet are used for extraction of semantics and sentiment polarity, with neural (CNN, RNN) and CRF [Lafferty et al. 2001] components for classification.

Furthermore, labelset-related ground truth has been explored for audio knowledge injection in NNs. For instance, ontology-aware approaches have utilized labelset class relationships—this is investigated in [Jiménez et al. 2018] in multiple ways: first, spectrograms are used with feed-forward NNs to directly model the ontology classes sequentially, i.e., reserving fully-connected layers to produce predictions for each ontology level. Each prediction is subsequently fed to the next level as input features. A second approach uses a siamese NN [Chicco 2021] trained on instance triplets with the Euclidean distance, with intra-class samples being encouraged to map to closely situated vectors in the embedding space. The network includes all potential cases in a two-layer ontology (i.e., matching subclass, only matching superclass, different superclass) and uses the same final classification method as the first architecture. Additional ontology-oriented studies include [Sun and Ghaffarzadegan 2020], where the authors consider labelset relations with a model consisting of a base CNN, followed by an LSTM with feed-forward and **graph convolutional networks** (**GCN**) [Zhang et al. 2019] modeling intra and interdependencies between levels of hierarchy in the ontology, respectively. Their system is evaluated in the single and multi-label classification of urban sounds. Additionally, the work in Zhang et al. [2021] utilizes the ASER eventuality knowledge graph [Zhang et al. 2020] to link acoustic event metadata descriptions with rich relations (e.g., conveying causality, temporal, and contingency relations). The generated associations are subsequently exploited via a relation-aware GCN variant for audio event categorization.

5 COMPARATIVE ANALYSIS

Having investigated characteristic methods of content-based and enrichment paradigms in the literature, we now move on to provide a critical comparison between them. We discuss common representation desiderata, summarized in a qualitative analysis in Table 4. Additionally, we provide indicative performance estimates in terms of classification accuracy;[1] these showcase a general trend of average performance and stability improvement as the complexity of a

[1] Reported performance stems from aggregating different experiments and datasets - see A.1 for a discussion on details, limitations and further results.

Table 4. Comparison between Content-based and Enrichment Paradigms

Desired Attributes	Content-based Paradigm			Enrichment Paradigm		
	LLTM	AGB	DRL	IEM	KBR	KAE
High-level semantics	X	?	✓	✓	✓	✓
Explainable	✓	?	X	✓	?	X
Data-driven/learned	X	?	✓	X	?	✓
Low-dimensional/space-efficient	?	✓	✓	X	✓	✓
Data efficient/lean	✓	✓	X	✓	✓	X
Computationally efficient	?	X	X	✓	?	X
Reusable/transferable	X	?	✓	✓	✓	✓
Avg. accuracy % (indicative)	86.2 ± 8.03	88.47 ± 7.98	90.78 ± 5.17	83.54 ± 13.34	84.78 ± 8.06	86.38 ± 7.27

Green checkmarks, red X's and yellow question marks indicate desired attributes that are true, false, or not determined / inconclusive, for the paradigm in the corresponding column. Paradigm average accuracy scores are approximated with the procedure detailed in the appendix A.1.

content-based/enrichment paradigm increases. At the same time, while enrichment efforts show promise, they are currently outperformed by large content-based classifiers engineered to score very large datasets. As the field of knowledge-augmented ML matures, we expect enrichment approaches to reach and/or surpass their content-based counterparts, along with providing additional benefits that come with structured knowledge (e.g., explainability).

5.1 Knowledge-agnostic Representation Paradigms

Section 2 covered broad approaches and indicative related work on content-based representations, organized in terms of the richness of semantics encapsulated in the output features. We now move on to a finer comparison, considering desired qualities for representations [Bengio et al. 2013] and general trends observed in the proposed paradigms.[2]

LLTM methods generally have no data needs outside the dataset of interest and create explainable representations: a feature coordinate has clear, non-ambiguous meaning, easily understood by referring to the generation algorithm. However, LLTM heavily relies on handcrafted feature engineering, demanding human expert intervention, knowledge and familiarity to the application domain. Features comparatively lack in richness of encapsulated semantics: very large vector spaces may be needed to arrive at adequate expressive power for efficient classification, while outputs are generally not reusable, leading to high-dimensional, specialized and often computationally demanding representations. As a result, cases with a severe lack of available data, or with high transparency/explainability requirements (e.g., medical/governance domains) could benefit from utilizing LLTM approaches.

AGB methods can generally build space-efficient representations, often configurable to a desired size for needs of specific tasks. AGB workflows use LLTM responses as inputs: as a result, they may require more data to reach meaningful results, along with increased need for computing power to fuel the additional processing steps. AGB generally produce distributed features, severely harming explainability. However, this improves the expressive power of the final representation and allows some degree of reusability—but no easy method for fine-tuning feature sets. Finally, such approaches may employ some degree of feature learning, but do so by using fixed, preconfigured rules and analytic solutions. Thus, AGB approaches could be favored if explainability requirements are low and not enough resources are available for paradigms of increased complexity and efficiency.

[2] While these observed trends generally hold and are indicative of paradigm approaches, edge cases, grey areas and exceptions are unavoidable.

Finally, DRL methods are fully data-driven, accumulating improvements in an incremental, partially stochastic manner. This generally produces semantically rich, distributed and compact features that are transferable and can be fine-tuned, but at the cost of creating black box-like feature extractors with low explainability regarding internal workings and output semantics—an issue which is an open and active area of research [Arras et al. 2017; Becker et al. 2018; Zhang and Zhu 2018]. Further, the reliance of these methods on distributional, data-driven operation renders them highly demanding with respect to the required amount of data and computational resources. As a result, deep representations could be the approach of choice if adequate compute and data resources can support their use, for tasks with very low explainability/transparency constraints.

Given the above, we observe that each approach has clear pros and cons with no definite, one-size-fits-all approach: the no free lunch theorem [Adam et al. 2019] appears to hold for representations approaches in classification. However, we propose that the historical progress of method evolution from low-level, to aggregation-based, to deep representations reflects research efforts to increase representation richness in semantics and conceptual information.

5.2 Representation Enrichment Approaches

We now move on to an analysis of representation enrichment paradigms and materials covered in Section 4. Overall, we can expect enrichment approaches to produce semantically rich features, provided an applicable knowledge resource of the desired domain and/or of sufficiently high-level information is utilized. Additionally, these paradigms generally present a high degree of compatibility with existing content-based features, i.e., can be utilized for transfer learning and feature reuse. Further, we provide knowledge-specific considerations for each enrichment paradigm below.

Input enrichment and modification (**IEM**) approaches require instance-level knowledge association: the resource must support mapping single, isolated instances to adequate (semantically and quantitatively meaningful) knowledge units. Thus, encompassed knowledge should meaningfully refer to singletons (e.g., E-ANEW word scores) and not be restricted to n-tuples (e.g., ParaphraseDB paraphrasal pairs). In the latter case, working solutions may be reachable by, e.g., partial/inverse mapping operations. Like the LLTM paradigm, IEM generally outputs enriched features without loss of explainability: content-based features are generally preserved, introducing knowledge-based information that is identifiable in the enriched representation and can be reviewed post-hoc. Since knowledge-based features constitute high-level conceptual information that can aid interpretation and provide intuition, this enrichment strategy could be favored by classification applications that prioritize explainability. The input space-oriented approach inherently supports reuse of content-based information—on the other hand, IEM's direct usage of knowledge features may make it sensitive to local outliers and redundancies (e.g., many-to-one mappings) introduced by the knowledge resource, giving rise to the need for filtering and/or normalization post-processing steps. Finally, IEM methods does not have large requirements in terms of data and compute requirements, but directly inflate the dimensionality of input features.

Knowledge-based refinement (**KBR**) exploits knowledge to guide enhancements and aggregations in groups of content-based representations. Contrary to IEM, this approach may require the knowledge resource to meaningfully refer to n-tuples of data (e.g., paraphrase pairs, semantic triples), linking them together via conceptual high-level relationships. This may restrict the number of resources usable with this enrichment scheme; however, desirable modifications may be viable with alignment operations [Amrouch and Mostefai 2012], data-driven methods, or by engaging with the internal organization to produce usable interfacing solutions, at the cost of suitable preprocessing steps. KBR aims at reusing content-based features by guided refinement, supporting transfer learning without data-driven fine-tuning, allowing pre-existing features to be repurposed for different tasks/domains, exploiting different knowledge resources as each use case demands.

However, careful configuration of the knowledge extraction process may be required to appropriately tune and regularize the contribution of different relevant components in the architecture, while the enriched representation may not be explainable (despite arising from explainable associations), as refining modifications may be applied in a distributed manner. Lastly, KBR enriches pre-existing features or generates knowledge-guided representations from scratch; as a result, the overall severity of its computational overhead is not clear.

Finally, **knowledge-aware end-to-end** (**KAE**) systems utilize knowledge in a holistic manner alongside content-based information, via deep representation learning. Knowledge utilization in KAE systems is versatile; instance/sub-instance knowledge unit (e.g., word/subword annotations in text) or n-tuple (e.g., ground truth tag/semantic relations) mappings may be used from a knowledge resource. Additionally, KAE shares attributes with its content-based counterpart (DRL), e.g., in terms of strong utilization of representation learning and representation dimensionality, but also with respect to the resulting enriched feature sets lacking explainability—i.e., relying on intuitions based on visualization / post-inspection techniques of model outputs and its internal representations [Mikolov et al. 2013; Yosinski et al. 2015]. Finally, training comes with requirements for large amounts of computational resources and data, which is only deteriorated by multiple resources and/or knowledge extraction and preprocessing steps that may be required.

In light of the above, in the next section, we discuss a high-level view of the material covered, along with findings, insights, and potential future directions of the representation enrichment field.

6 DISCUSSION AND RESEARCH FINDINGS

In this section, we provide a high-level view of the totality of the covered material, organized in a set of research questions and findings. From a study of the related literature, we can recognize the following:

(a) Knowledge-based enrichment offers a route for enhancing explainability. In Machine Learning there exists a duality between learning and forming a representation: finding good representations for a given task implies facilitating learning. This has been further accentuated through deep learning approaches [Adadi and Berrada 2018; Burkart and Huber 2021] and emphasized in this work. Given this relation, one would expect that explainability requirements in the learning process could also directly address the representations itself. As argued in this study, enriching representations provides content-agnostic avenues for improving representation explainability, via enriched features that are grounded in—or directly contain—explainable knowledge. This potential seems especially promising in deep neural features, where the performance/explainability tradeoff is most severe.

(b) Research trends towards Representation Learning. The proposed content-based categories provide a view that reflects a general historical evolution of main research efforts in data representations and resulting increase in representation complexity and richness of encapsulated semantics: simpler LLTM-based features are becoming increasingly inadequate, while static aggregation/transformation employed by AGB is lacking. Instead, research has shifted to favor strategies with less inductive biases, like DRL. Notably, this shift appears to carry over to representation enrichment; observed trends show an increase in complexity, stochasticity and computational cost of knowledge utilization. Namely, IEM performs simple operations on the feature set of the content-based baseline (e.g., expansion), while KBR applies well-defined engineered modifications at predefined points in the computation, when/where knowledge application is deemed relevant. On the other hand, the enrichment mechanism in KAE systems is entirely delegated to data-driven learning that jointly leverages content and knowledge information, an approach favored in the enrichment domain, as reflected in our literature review and indicated by publication trends (e.g., appendix Figure A.1).

(c) Knowledge-based enrichment presents a promising alternative for rich representations. As stressed in Section 2, content-based paradigms have pros and cons with no one-size-fits-all solution. In light of this, knowledge-based enrichment remains a promising alternative for arriving at rich features, without paying the cost of AGB/DRL (e.g., over-engineering, large complexity jumps and data/compute requirements), while offering control of what knowledge is infused (e.g., pertaining to domains of interest), on which content-based features, and how this enrichment is applied.

(d) Further research is required for optimal knowledge-based representation enrichment. Current enrichment approaches are influenced by content-based paradigms, carrying over working solutions but perpetuating disadvantages. This makes selecting an overall optimal approach difficult, hence the fine-grained comparisons/use-case-specific suggestions provided in Section 5.2. Paradigm-specific constraints imply that careful selection of the resource/enrichment method may be necessary, on a per-application basis. Further research is required to arrive at robust methods, highly compatible with different types of instances. This is necessary, given that knowledge appears to still play an auxiliary/complementary role to content [Pittaras et al. 2020], a remark compounded by modality-centric issues highlighted in subsequent findings. However, novel techniques are continuously being invented, as exemplified by the successful integration of enrichment with deep learning methods in KAE, enabling the exploitation of knowledge in representation learning in a holistic manner, via leveraging both content-based and high-quality structured information in state-of-the-art classification pipelines. We believe that KAE shows promise in the search of optimally fusing distributional and knowledge-oriented information, exploiting the best of both worlds.

(e) Semantic delineation affects representations, knowledge compatibility and enrichment. In this survey we covered works handling text, image and audio data modalities, associated with qualitatively different semantic gaps: for text, high-quality semantic segmentation is readily available by virtue of existing rules in language, syntax and grammar. In contrast, such rules for multimedia are hidden away in our visual/auditory systems; representations for such data are limited to operations on signal values with little to no direct linkage to high-level information [Sethi et al. 2001]. As observed in Section 2, this shapes the description, localization and extraction of content-based representations. Moreover, these differences may render compiling complex knowledge for non-textual modalities difficult, leading enrichment techniques to largely rely on knowledge of linguistic nature, that has the capacity to encapsulate higher-level semantics. Indeed, many multimedia-related knowledge resorts to metadata annotations on large datasets, following a more data-driven approach. This has limited the enrichment of image/audio to indirect knowledge utilization—i.e., via exploiting tags and labelset structure through their textual descriptions and lexicalizations. To this end, producing knowledge resources suitable for direct exploitation in image and audio representation enrichment would be a step towards additional improvements on the performance and interpretation of image/audio classification models.

7 CONCLUSION

In this survey, we investigated representation enrichment approaches from external knowledge resources, covering different modalities (text, images, and audio) and feature generation techniques, in the context of classification. Related literature was organized in distinctive categories with respect to the representation paradigm employed and organized in summary tables to facilitate comparison and lookup. We began by cataloguing knowledge-agnostic representations, organized in three broad categories, namely (a) low-level/template-matching methods, that extract low-level information and handcrafted features (b) aggregation-based approaches, that combine, transform or post-process low-level results, and (c) deep representation systems, that build hierarchical feature sets via deep representation learning. Qualitative comparisons and use cases suggestions are provided for each paradigm.

We moved on to expand on the motivation for utilizing knowledge in classification, listing available exploitable resources, along with details regarding their information content, structure, and retrieval details. Enriched representations are covered next, i.e., studies that take advantage of such knowledge resources, grouping related work into groups of similar enrichment methodology, namely (a) input enrichment and modification, covering works that inject knowledge-based information in the input feature space, (b) knowledge-based refinement, which aggregates and/or transforms representations via knowledge-determined operations, and (c) knowledge-aware end-to-end systems, consisting of pipelines that jointly learn nonlinear feature hierarchies from content and knowledge inputs. Finally, we compare enrichment categories in representation and knowledge-oriented contexts and discuss research findings.

There are multiple ways to complement/extend the work in this survey. Our primary focus was the representation component of the classification pipeline—one avenue for future work would be the investigation of enrichment approaches that focus on the learning algorithm, i.e., independent of the representation approach. Additionally, an exploration of knowledge utilization between classification and other ML tasks (e.g., in a comparative study) or a task-agnostic approach would be beneficial towards a better understanding of the wider impact of enrichment. Further, meta-analysis on selected works with comparable experimental setups could be conducted to quantitatively assess the impact of different representation and enrichment paradigms. Finally, the representation building and enrichment paradigms proposed in this survey could be investigated in the context of the classification of multimodal instances.

A APPENDIX

A.1 Performance Comparison

Here we describe the approach for computing the aggregated performance scores for the proposed paradigms, as referenced in the table and relevant discussion in Section 5. Given that in this study we cover works that tackle different classification tasks and modalities, there is a great variability in benchmark datasets, dataset versions, slices and subsets, specific classification subtasks and evaluation metrics used in each related work item examined. This makes the exact comparison of approaches very difficult, since it has not been possible to determine a shared setting (e.g., common dataset and metric) for each modality in question. As a result, we are forced to group studies that use different evaluation protocols—we would thus like to stress that the outcomes of this comparison, while useful, should not be considered as concrete, definite evidence, but serve instead as general hints and noisy indications of performance trends for each representation paradigm.

We adopt accuracy as the evaluation metric, being the most prolific performance measure in the related work we discuss in the survey. For studies that evaluate multiple datasets, we average the top 3 results in terms of test set performance of the proposed approach. We collect accuracy scores, dataset/labelset sizes and number of datasets utilized in the evaluation. We average this information for each content-based and enrichment paradigm, using the two most recent studies from the related work discussed in Sections 2 and 4 that report accuracy-based classification. Results are thus computed by aggregating 6 articles per paradigm, with a pool of 36 articles used to produce the entire body of reported results. This information is summarized with mean and standard deviation values in Table A.1.

Along with trends for average accuracy appearing to increase as paradigms become more complex (discussed in the beginning of Section 5), we provide per-category scores for the number of datasets utilized for evaluation and their instance/labelset sizes. These also showcase upward

Table A.1. Classification Performance (Accuracy %) and Dataset-related Statistics (Count, Sample / Labelset Size) of Articles Discussed in This Study, Averaged by Content-based and Knowledge-based Enrichment Paradigms Proposed in the Survey

Content-based Paradigms			
Statistic	LLTM	AGB	DRL
Accuracy %	86.2 ± 8.03	88.47 ± 7.98	90.78 ± 5.17
Dataset size	626.08 ± 427.76	55, 665.85 ± 116, 282.2	2, 279, 126.92 ± 2, 862, 928.64
Number of datasets	1.33 ± 0.52	5.5 ± 5.43	2.83 ± 3.54
Labelset size	4.67 ± 3.01	43.82 ± 87.9	2, 104.18 ± 4619.3

Enrichment Paradigms			
Statistic	IEM	KBR	KAE
Accuracy %	83.54 ± 13.34	84.78 ± 8.06	86.38 ± 7.27
Dataset size	3, 108.02 ± 3433.92	242, 867.72 ± 533, 991.79	80, 585.86 ± 80, 412.9
Number of datasets	3.17 ± 5.31	2.17 ± 1.33	2.67 ± 1.75
Labelset size	5.63 ± 2.87	420.56 ± 415.23	231.58 ± 266.11

See A.1 for a discussion on details and limitations of the approach by which these values where computed.

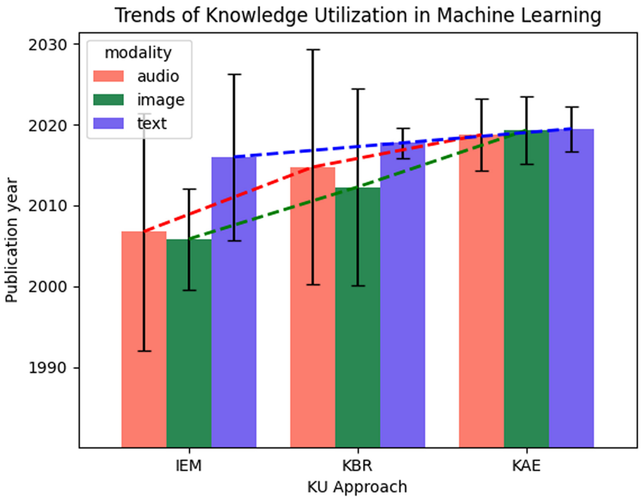

Fig. A.1. Mean publication year for ≈50 knowledge-enrichment works considered, grouped by knowledge enrichment paradigm and modality. We can observe an upwards trend towards deep representation learning for knowledge utilization, reflecting the historical evolution of enrichment approaches.

trends in accordance with approach complexity for content-based methods (i.e., from LLTM to AGB to DRL); on the other hand, these statistics do not seem to follow such a monotonic trend for enrichment approaches. This could be explained by the fact that enrichment workflows include external knowledge resources in their processing and learning pipelines, the utilization of which often results in significant expansion of input information, and thus of the effective input dataset. This, along with additional computational cost requirements from knowledge integration in the learning method, could act as considerable limiting factors in scaling dataset counts and sizes when adopting representation strategies of greater complexity; this could in turn explain the more conservative experiment scaling in IEM, KBR, and KAE paradigms, compared to their content-based counterparts.

To conclude, we offer this brief comparative snapshot along with the larger, qualitative discussion in Section 5 in order to provide readers with an approximate view of how content-based and enrichment methods stand, with respect to classification performance and scale of empirical analysis. A rigorous quantitative investigation (e.g., application of selected systems to common, fixed benchmarks and evaluation protocols) is out of the scope of this survey, and we reserve such efforts for future work.

REFERENCES

A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

Stavros P. Adam, Stamatios-Aggelos N. Alexandropoulos, Panos M. Pardalos, and Michael N. Vrahatis. 2019. No free lunch theorem: A review. *Approximation and Optimization* 145 (2019), 57–82.

C. C. Aggarwal. 2015. Data classification. In *Proceedings of the Data Mining*. Springer, 285–344.

P. F. Alcantarilla, A. Bartoli, and A. J. Davison. 2012. KAZE features. In *Proceedings of the European Conference on Computer Vision*. Springer, 214–227.

Alo Allik, György Fazekas, and Mark B. Sandler. 2016. An ontology for audio features. In *Proceedings of the ISMIR*. 73–79.

B. Altınel and M. C. Ganiz. 2018. Semantic text classification: A survey of past and recent advances. *Information Processing and Management* 54, 6 (2018), 1129–1153.

Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. 2015. Fast image classification for monument recognition. *Journal on Computing and Cultural Heritage* 8, 4 (2015), 1–25.

Siham Amrouch and Sihem Mostefai. 2012. Survey on the literature of ontology mapping, alignment and merging. In *Proceedings of the 2012 International Conference on Information Technology and e-Services*. IEEE, 1–5.

L. H. Anaya. 2011. *Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers*. ERIC.

X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 2 (2012), 356–370.

E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran. 2012. Deep neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. 20–28.

L. Arras, F. Horn, G. Montavon, K. Müller, and W. Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PloS One* 12, 8 (2017), e0181142.

N. Aye, F. Hattori, and K. Kuwabara. 2008. Use of ontologies for bridging semantic gaps in distant communication. *International Conference on Innovations in Information Technology* (2008), 371–375.

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Lrec*, Vol. 10. 2200–2204.

Dima Badawi and Hakan Altınçay. 2014. A novel framework for termset selection and weighting in binary text classification. *Engineering Applications of Artificial Intelligence* 35 (2014), 38–53.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR'15)*.

M. R. Bai and M. Chen. 2007. Intelligent preprocessing and classification of audio signals. *Journal of the Audio Engineering Society* 55, 5 (2007), 372–384.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, Morgan Kaufmann Publishers / ACL, 86–90.

B. K. Baniya, J. Lee, and Z. Li. 2014. Audio feature reduction and analysis for automatic music genre classification. In *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 457–462.

M. Baroni, G. Dinu, and G. Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 238–247.

M. Basseville. 1989. Distance measures for signal processing and pattern recognition. *Signal Processing* 18, 4 (1989), 349–369.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110, 3 (2008), 346–359.

Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2018. Interpreting and Explaining Deep Neural Networks for Classification of audio signals. CoRR abs/1807.03418.

R. Bellman. 2013. *Dynamic Programming*. Courier Corporation.

Y. Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends®in Machine Learning* 2, 1 (2009), 1–127. DOI: https://doi.org/10.1561/2200000006

Y. Bengio. 2011. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the JMLR Work*. JMLR.org, 1–20. DOI: https://doi.org/10.1109/IJCNN.2011.6033302

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

Y. Bengio, O. Delalleau, and N. Le Roux. 2005. The curse of dimensionality for local kernel machines. *Techn. Rep.* 1258 (2005), 12.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. 2007. Greedy layer-wise training of deep networks. In *Proceedings of the Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA., 153–160.

A. B. Benitez and S. Chang. 2003. Image classification using multimedia knowledge networks. In *Proceedings of the 2003 International Conference on Image Processing*. IEEE, III–613.

D. Bertero and P. Fung. 2016. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 496–501.

S. Bhattacharyya. 2011. A brief survey of color image preprocessing and segmentation techniques. *JPRR* 6, 1 (2011), 120–129. DOI: https://doi.org/10.13176/11.191

X. Bian, C. Chen, L. Tian, and Q. Du. 2017. Fusing local and global features for high-resolution scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10, 6 (2017), 2889–2901.

A. Binder, M. Kawanabe, and U. Brefeld. 2009. Efficient classification of images with taxonomies. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 351–362.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

Jonathan Bodine and Dorit S. Hochbaum. 2022. A better decision tree: The max-cut decision tree with modified PCA improves accuracy and running time. *SN Computer Science* 3, 4 (2022), 1–18.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. ACM, 1247–1250.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the Advances in Neural Information Processing Systems*. 2787–2795.

A. Borghesi, F. Baldo, and M. Milano. 2020. Improving deep learning models via constraint-based domain knowledge: A brief survey. arXiv:2005.10691. Retrieved from https://arxiv.org/abs/2005.10691.

H. Boyer, X. Serra, and G. Peeters. 1999. Audio descriptors and descriptor schemes in the context of MPEG-7. In *Proceedings of the 1999 International Computer Music Conference*. International Computer Music Conference.

R. N. Bracewell and R. N. Bracewell. 1986. *The Fourier Transform and its Applications*. Vol. 31999. McGraw-Hill New York.

M. M. Bradley and P. J. Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Technical Report. Technical report C-1, the center for research in psychophysiology.

C. J. Burges, J. C. Platt, and S. Jana. 2002. Extracting noise-robust features from audio data. In *Proceedings of the s2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, I–1021.

N. Burkart and M. F. Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

M. Cadoli and F. M. Donini. 1997. A survey on knowledge compilation. *AI Communications* 10, 3–4 (1997), 137–150. Retrieved from http://content.iospress.com/articles/ai-communications/aic133.

M. Calonder, V. Lepetit, C. Strecha, and P. Fua. 2010. Brief: Binary robust independent elementary features. In *Proceedings of the European Conference on Computer Vision*. Springer, 778–792.

J. Camacho-Collados and M. T. Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research* 63 (2018), 743–788.

F. Camastra and A. Vinciarelli. 2015. *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. Springer.

P. Cano, M. Koppenberger, P. Herrera, S. Le Groux, J. Ricard, and N. Wack. 2004. Nearest-neighbor generic sound classification with a WordNet-based taxonomy. In *Proceedings of the Audio Engineering Society Convention 116*. Audio Engineering Society.

S. Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1, 2 (2007), 1.

Simyung Chang, Hyoungwoo Park, Janghoon Cho, Hyunsin Park, Sungrack Yun, and Kyuwoong Hwang. 2021. Subspectral normalization for neural audio data processing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 850–854.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 357–366.

Junyi Chen, Shankai Yan, and Ka-Chun Wong. 2020. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications* 32, 15 (2020), 10809–10818.

Q. Chen, X. Zhu, Z. Ling, D. Inkpen, and S. Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2406–2417.

Heng-Tze Cheng, Yi-Hsuan Yang, Yu-Ching Lin, I-Bin Liao, and Homer H. Chen. 2008. Automatic chord recognition for music classification and retrieval. In *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo*. IEEE, 1505–1508.

D. Chicco. 2021. Siamese neural networks: An overview. *Artificial Neural Networks* 2190 (2021), 73–94.

K. Choi, G. Fazekas, and M. Sandler. 2016. Explaining deep convolutional neural networks on music classification. CoRR abs/1607.02444.

X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*. 2201–2206.

A. M. Collins and E. F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82, 6 (1975), 407.

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 886–893.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 447–459.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407.

J. Delhumeau, P. Gosselin, H. Jégou, and P. Pérez. 2013. Revisiting the VLAD image representation. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 653–656.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Computer Vision and Pattern Recognition*. IEEE, 248–255.

P. Deng. 1990. Inducing decision-making knowledge from data bases: An approach to automating knowledge acquisition. In *Proceedings of the 1990 ACM SIGBDP Conference on Trends and Directions in Expert Systems*. Elias M. Awad (Ed.), ACM, 189–211. DOI: https://doi.org/10.1145/97709.97725

T. Deselaers and V. Ferrari. 2011. Visual and semantic similarity in ImageNet. In *Proceedings of the CVPR*. IEEE Computer Society, 1777–1784. Retrieved from http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#DeselaersF11.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning*. PMLR, 647–655.

F. K. Došilović, M. Brčić, and N. Hlupić. 2018. Explainable artificial intelligence: A survey. In *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics*. IEEE, 0210–0215.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Z. Elberrichi, A. Rahmoun, and M. A. Bentaalah. 2008. Using WordNet for text categorization. *International Arab Journal of Information Technology* 5, 1 (2008).

M. Everingham, A. Zisserman, C. K. Williams, L. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, and G. Dorkó. 2005. The 2005 pascal visual object classes challenge. In *Proceedings of the Machine Learning Challenges Workshop*. Springer, 117–176.

Xavier Favory, Konstantinos Drossos, Tuomas Virtanen, and Xavier Serra. 2020. COALA: Co-aligned autoencoders for learning semantically enriched audio representations. In *Proceedings of the International Conference on Machine Learning*.

P. Ferragina and U. Scaiella. 2010. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. 1625–1628.

L. Ferrone and F. M. Zanzotto. 2020. Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. *Frontiers in Robotics and AI* 6 (2020), 153.

C. J. Fillmore. 2006. Frame semantics. *Cognitive Linguistics: Basic Readings* 34 (2006), 373–400.

C. J. Fillmore, C. R. Johnson, and M. R. L. Petruck. 2003. Background to framenet. *International Journal of Lexicography* 16, 3 (2003), 235–250. DOI: http://dx.doi.org/10.1093/ijl/16.3.235

Z. Fu, G. Lu, K. M. Ting, and D. Zhang. 2010. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia* 13, 2 (2010), 303–319.

J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 758–764.

S. García, J. Luengo, and F. Herrera. 2015. *Data Preprocessing in Data Mining.* Vol. 72. Springer.

J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE ICASSP 2017.* New Orleans, LA.

A. Gersho and R. M. Gray. 1992. *Vector Quantization and Signal Compression.* Vector Quantization and Signal Compression. DOI: https://doi.org/10.1007/978-1-4615-3626-0

G. Giannakopoulos, P. Mavridi, G. Paliouras, G. Papadakis, and K. Tserpes. 2012. Representation models for text classification: A comparative analysis over three web document types. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics.* ACM, 13.

G. Glavaš and I. Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 34–45.

J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov. 2004. Neighbourhood components analysis. *Advances in Neural Information Processing Systems* 17 (2004), 513–520.

Gene H. Golub. 1969. Matrix decompositions and statistical calculations. In *Proceedings of the Statistical Computation.* Elsevier, 365–397.

Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: Audio spectrogram transformer. In *Interspeech 2021, 22nd Annual Conference of the Inter National Speech Communication Association (ISCA'21 Brno, Czechia, 30 August-3 September 2021)*, 571–575.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR'15, San Diego, CA, USA, May 7-9, 2015)*, Conference Track Proceedings.

Roger B. Grosse, Rajat Raina, Helen Kwong, and Andrew Y. Ng. 2007. Shift-Invariance Sparse Coding for Audio Classification. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI'07, Vancouver, BC, Canada, July 19-22, 2007)*, AUAI Press, 149–158.

J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, and J. Cai. 2018. Recent advances in convolutional neural networks. *Pattern Recognition* 77 (2018), 354–377.

Z. S. Harris. 1954. Distributional structure. *Word* 10, 2–3 (1954), 146–162.

Guiqing He, Feng Li, Qiyao Wang, Zongwen Bai, and Yuelei Xu. 2021. A hierarchical sampling based triplet network for fine-grained image classification. *Pattern Recognition* 115 (2021), 107889.

H. He and Y. Ma. 2013. Imbalanced learning: Foundations, algorithms, and applications. Wiley-IEEE Press.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, and B. Seybold. 2017. CNN architectures for large-scale audio classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 131–135.

G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. 1984. *Distributed Representations.* Carnegie-Mellon University Pittsburgh, PA.

S. Hochreiter and J. J. Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1–32. https://doi.org/10.1162/neco.1997.9.8.1735

X. Hu and J. S. Downie. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries.* 159–168.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4700–4708.

D. H. Hubel and T. N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology* 160, 1 (1962), 106–154.

Ioana Ilea, Lionel Bombrun, Christian Germain, Romulus Terebes, Monica Borda, and Yannick Berthoumieu. 2016. Texture image classification with Riemannian Fisher vectors. In *Proceedings of the 2016 IEEE International Conference on Image Processing.* IEEE, 3543–3547.

A. K. Jain and R. C. Dubes. 1988. *Algorithms for Clustering Data.* Prentice-Hall.

Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. 2015. Emotion analysis of songs based on lyrical and audio features. CoRR abs/1506.05012 (2015).

M. Jarmasz and S. Szpakowicz. 2004. Roget's thesaurus and semantic similarity. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP* 2003 (2004), 111.

Mirantha Jayathilaka, Tingting Mu, and Uli Sattler. 2021. Ontology-based n-ball Concept Embeddings Informing Few-shot Image Classification. In *Machine Learning with Symbolic Methods and Knowledge Graphs co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021), Virtual, September 17, 2021 (CEUR Workshop Proceedings)*, Vol. 2997. CEUR-WS.org.

A. Jiménez, B. Elizalde, and B. Raj. 2018. Sound event classification using ontology-based neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (2019), 3365–3385.

A. G. Jivani. 2011. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl.* 2, 6 (2011), 1930–1938. Retrieved from https://www.researchgate.net/profile/Anjali.

J. Jolion and W. Kropatsch. 2012. *Graph based Representations in Pattern Recognition*. Vol. 12. Springer Science & Business Media.

I. Jolliffe. 2011. Principal component analysis. In *Proceedings of the International Encyclopedia of Statistical Science*. Springer, 1094–1096.

C. Jörgensen, A. Jaimes, A. B. Benitez, and S. Chang. 2001. A conceptual framework and empirical research for classifying visual descriptors. *Journal of the American Society for Information Science and Technology* 52, 11 (2001), 938–947.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, (EACL'17, Valencia, Spain, April 3-7, 2017)* Volume 2: Short Papers. Association for Computational Linguistics, 427–431.

J. S. Justeson and S. M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1, 1 (1995), 9–27.

P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online, 6975–6988. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.567

J. Kennedy and R. Eberhart. 1995. Particle swarm optimization. In *Proceedings of the ICNN'95-International Conference on Neural Networks*, Vol. 4. IEEE, 1942–1948.

S. Kim, P. Georgiou, and S. Narayanan. 2012. Latent acoustic topic models for unstructured audio classification. *APSIPA Transactions on Signal and Information Processing* 1 (2012), e6.

T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo. 2008. Combining image captions and visual analysis for image concept classification. In *Proceedings of the 9th International Workshop on Multimedia Data Mining: Held in conjunction with the ACM SIGKDD 2008*. 8–17.

E. C. Knight, S. Poo Hernandez, E. M. Bayne, V. Bulitko, and B. V. Tucker. 2020. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics* 29, 3 (2020), 337–355.

T. Kohonen. 1990. The self-organizing map. *Proc. IEEE* 78, 9 (1990), 1464–1480.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, (CoNLL'18, Brussels, Belgium, October 31 - November 1, 2018)*, Association for Computational Linguistics, 519–529.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* 123, 1 (2017), 32–73.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Adv. Neural Inf. Process. Syst.* Curran Associates, Inc., 1097–1105. Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

V. Kumar and S. Minz. 2013. Mood classifiaction of lyrics using SentiWordNet. In *Proceedings of the 2013 International Conference on Computer Communication and Informatics*. IEEE, 1–5.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01, Williams College, Williamstown, MA, USA, June 28 - July 1, 2001)*, Morgan Kaufmann, 282–289.

C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen. 2009. Exploring relationships between audio features and emotion in music. In *Proceedings of the ESCOM 2009: 7th Triennial Conference of European Society for the Cognitive Sciences of Music*.

S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 2169–2178.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. 2007. Efficient sparse coding algorithms. In *Proceedings of the Proceedings of the Advances in Neural Information Processing Systems*. 801–808.

K. Lee and D. P. Ellis. 2010. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 6 (2010), 1406–1416.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17, Copenhagen, Denmark, September 9-11, 2017)*. Association for Computational Linguistics, 188–197.

J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, and S. Auer. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

D. B. Lenat, R. V. Guha, K. Pittman, D. Pratt, and M. Shepherd. 1990. Cyc: Toward programs with common sense. *Communications of the ACM* 33, 8 (1990), 30–49.

Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. 2011. BRISK: Binary robust invariant scalable keypoints. In *Proceedings of the 2011 International Conference on Computer Vision*. Ieee, 2548–2555.

D. D. Lewis and M. Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, Vol. 33. 81–93.

Guohao Li, Xin Wang, and Wenwu Zhu. 2019. Perceptual visual reasoning with knowledge propagation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 530–538.

Tian Li, Xiang Chen, Zhen Dong, Kurt Keutzer, and Shanghang Zhang. 2022. Domain-adaptive text classification with Structured Knowledge from Unlabeled Data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI'22)*. Vienna, Austria, ijcai.org, 4216–4222.

W. Li, L. Niu, and D. Xu. 2014. Exploiting privileged information from web data for image categorization. In *Proceedings of the European Conference on Computer Vision*. Springer, 437–452.

W. Li and M. Sun. 2006. Automatic image annotation based on WordNet and hierarchical ensembles. In *Proceedings of the CICLing*. Alexander F. Gelbukh (Ed.), Lecture Notes in Computer Science, Vol. 3878, Springer, 417–428. Retrieved from http://dblp.uni-trier.de/db/conf/cicling/cicling2006.html#LiM06.

Y. Li, Q. Pan, T. Yang, S. Wang, J. Tang, and E. Cambria. 2017. Learning word representations for sentiment analysis. *Cognitive Computation* 9, 6 (2017), 843–851.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated graph sequence neural networks. In *4th International Conference on Learning Representations (ICLR'16, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings)*.

Y. H. Li and A. K. Jain. 1998. Classification of text documents. *The Computer Journal* 41, 8 (1998), 537–546.

H. Liu and P. Singh. 2004. ConceptNet – a practical commonsense reasoning toolkit. *BT Technology Journal* 22, 4 (2004), 211–226.

Yingying Liu, Peipei Li, and Xuegang Hu. 2022. Combining context-relevant features with multi-stage attention network for short text classification. *Computer Speech and Language* 71 (2022), 101268.

Y. Liu, Z. Liu, T. Chua, and M. Sun. 2015. Topical word embeddings. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. Citeseer.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

D. G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

D. Lu and Q. Weng. 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28, 5 (2007), 823–870.

J. Luo and A. Savakis. 2001. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In *Proceedings of the 2001 International Conference on Image Processing*. IEEE, 745–748.

B. S. Manjunath and W. Ma. 1996. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18, 8 (1996), 837–842.

C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Scoring, Term Weighting, and the Vector Space Model*. Cambridge University Press, 100–123. DOI: https://doi.org/10.1017/CBO9780511809071.007

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2017. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society)*. 20–28.

Ladislav Maršík, J. Pokornyy, and Martin Ilcík. 2014. Improving music classification using harmonic complexity. In *Proceedings of the 14th Conference Information Technologies-Applications and Theory*. 13–17.

M. Marszalek and C. Schmid. 2007. Semantic hierarchies for visual object recognition. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–7.

S. Martinčić-Ipšić, T. Miličić, and L. Todorovski. 2019. The influence of feature representation of text on the performance of document classification. *Applied Sciences* 9, 4 (2019), 743.

L. R. Medsker and L. Jain. 2001. Recurrent neural networks. *Design and Applications* 5 (2001).

R. Mehrotra, K. R. Namuduri, and N. Ranganathan. 1992. Gabor filter-based edge detection. *Pattern Recognition* 25, 12 (1992), 1479–1494.

Julia A. Meister, Khuong An Nguyen, and Zhiyuan Luo. 2022. Audio feature ranking for sound-based COVID-19 patient detection. In *Progress in Artificial Intelligence - 21st EPIA Conference on Artificial Intelligence (EPIA'22, Lisbon, Portugal, August 31 - September 2, 2022, Proceedings) (Lecture Notes in Computer Science)*, Vol. 13566. Springer, 146–58.

Cui Menglong, Ji Detao, Zeng Ting, Zhang Dehai, Xie Cheng, Chen Zhibo, and Xia Xiaoqiang. 2019. Image classification based on image knowledge graph and semantics. In *Proceedings of the 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design*. IEEE, 81–86.

N. Mesgarani, M. Slaney, and S. A. Shamma. 2006. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 3 (2006), 920–930.

P. Miettinen. 2009. Matrix decomposition methods for data mining: Computational complexity and algorithms. (2009).

K. Mikolajczyk, B. Leibe, and B. Schiele. 2005. Local features for object class recognition. In *Proceedings of the 10th IEEE International Conference on Computer Vision*. IEEE, 1792–1799.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*. Curran Associates, Inc., 3111–3119.

G. A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, 11 (1995), 39–41.

S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys* 54, 3 (2021), 40 pages. DOI: https://doi.org/10.1145/3439726

M. Mohri, A. Rostamizadeh, and A. Talwalkar. 2018. *Foundations of Machine Learning*. MIT press.

Seyyed Hamid Samareh Moosavi and Vahid Khatibi Bardsiri. 2019. Poor and rich optimization algorithm: A new human-based and multi populations algorithm. *Engineering Applications of Artificial Intelligence* 86 (2019), 165–181.

Loris Nanni, Gianluca Maguolo, Sheryl Brahnam, and Michelangelo Paci. 2021. An ensemble of convolutional neural networks for audio classification. *Applied Sciences* 11, 13 (2021), 5796. DOI: https://doi.org/10.3390/app11135796

R. Navigli and S. P. Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 216–225.

H. Nezreg, H. Lehbab, and H. Belbachir. 2014. Conceptual representation using wordnet for text categorization. *International Journal of Computer and Communication Engineering* 3, 1 (2014), 27.

A. D. Ningtyas, E. B. Nababan, and S. Efendi. 2022. Performance analysis of local binary pattern and k-nearest neighbor on image classification of fingers leaves. *International Journal of Nonlinear Analysis and Applications* 13, 1 (2022), 1701–1708.

Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, and Bohyung Han. 2019. Transfer learning via unsupervised task discovery for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8385–8394.

T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 971–987.

A. Oliva and A. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

A. Oliva and A. Torralba. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research* 155 (2006), 23–36.

Bruno A. Olshausen and David J. Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 6583 (1996), 607–609.

F. Pachet and P. Roy. 2009. Analytical features: A knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing* 2009 (2009), 1–23.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11130–11138.

S. Pei and C. Lin. 1995. Image normalization for pattern recognition. *Image and Vision computing* 13, 10 (1995), 711–723.

J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

F. Perronnin and C. Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT'18, New Orleans, Louisiana, USA, June 1-6, 2018) Volume 1 (Long Papers), Association for Computational Linguistics, 2227–2237.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, (EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019)*, Association for Computational Linguistics, 43–54.

N. Pittaras, G. Giannakopoulos, G. Papadakis, and V. Karkaletsis. 2020. Text classification with semantically enriched word embeddings. *Natural Language Engineering* 27, 4 (2020), 1–35. DOI: https://doi.org/10.1017/S1351324920000170

N. Pittaras, F. Markatopoulou, V. Mezaris, and I. Patras. 2017. Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 102–114.

S. Anuja Prasad and Leena Mary. 2019. A comparative study of different features for vehicle classification. In *Proceedings of the 2019 International Conference on Computational Intelligence in Data Science*. IEEE, 1–5.

Y. Raimond, S. A. Abdallah, M. B. Sandler, and F. Giasson. 2007. The music ontology. In *Proceedings of the ISMIR*. Citeseer, 8th.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conf. Empir. Methods Nat. Lang. Process.*

M. I. Razzak, S. Naz, and A. Zaib. 2018. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps* (2018), 323–350.

Douglas A. Reynolds. 2009. Gaussian Mixture Models. In *Encyclopedia of Biometrics*. Springer, 659–663.

Seungmin Rho, Seheon Song, Eenjun Hwang, and Minkoo Kim. 2009. COMUS: Ontological and rule-based reasoning for music recommendation system. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 859–866.

Vladimir Risojević, Snježana Momić, and Zdenka Babić. 2011. Gabor descriptors for aerial image classification. In *Proceedings of the International Conference on Adaptive and Natural Computing Algorithms*. Springer, 51–60.

L. Rokach and O. Maimon. 2005. Clustering methods. In *Proceedings of the Data Mining and Knowledge Discovery Handbook*. Springer, 321–352.

E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the 2011 International Conference on Computer Vision*. IEEE, 2564–2571.

D. Rumelhart, G. Hinton, and R. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323 (1986), 533–536.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (1988), 513–523.

G. Salton, A. Wong, and C. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18, 11 (1975), 613–620.

H. Schmid. 1994. TreeTagger-a language independent part-of-speech tagger. (1994). Retrieved from http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1 (2002), 1–47.

E. Sejdić, I. Djurović, and J. Jiang. 2009. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing* 19, 1 (2009), 153–183.

I. K. Sethi, I. L. Coman, and D. Stan. 2001. Mining association rules between low-level image features and high-level concepts. In *Proceedings of the Data Mining and Knowledge Discovery: Theory, Tools, and Technology III*, Vol. 4384. International Society for Optics and Photonics, 279–290.

Weijia Shi, Muhao Chen, Pei Zhou, and Kai-Wei Chang. 2019. Retrofitting Contextualized Word Embeddings with Paraphrases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP'19, Hong Kong, China, November 3-7, 2019)*, Association for Computational Linguistics, 1198–1203.

C. Shorten and T. M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. 6 (2019). DOI: https://doi.org/10.1186/s40537-019-0197-0

Leslie F. Sikos. 2017. The Semantic Gap. *Description Logics in Multimedia Reasoning* (2017), 51–66.

C. N. Silla and A. A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22, 1 (2011), 31–72.

C. Silva and B. Ribeiro. 2003. The importance of stop word removal on recall values in text categorization. In *Proceedings of the Neural Networks, 2003. Proc. Int. Jt. Conf.*, Vol. 3. IEEE, 1661–1666.

Mattia Silvestri, Michele Lombardi, and Michela Milano. 2021. Injecting domain knowledge in neural networks: a controlled experiment on a constrained problem. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research: 18th International Conference (CPAIOR'21, Vienna, Austria, July 5–8, 2021, Proceedings 18)*. Springer, 266–282.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR'15. San Diego, CA, USA, May 7-9, 2015)*, Conference Track Proceedings.

S. Singh. 2013. Optical character recognition techniques: A survey. *Journal of Emerging Trends in Computing and Information Sciences* 4, 6 (2013), 545–550.

B. Škrlj, M. Martinc, J. Kralj, N. Lavrač, and S. Pollak. 2020. tax2vec: Constructing interpretable features from taxonomies for short text classification. *Computer Speech and Language* 65 (2020), 101104.

Blaž Škrlj, Matej Martinc, Nada Lavrač, and Senja Pollak. 2021. autoBOT: Evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning* 110, 5 (2021), 989–1028.

M. Slaney. 1998. Auditory toolbox. *Interval Research Corporation, Tech. Rep.* 10, 1998 (1998).

M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45, 4 (2009), 427–437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002

S. S. Sonawane and P. A. Kulkarni. 2014. Graph based representation and analysis of text document: A survey of techniques. *International Journal of Computer Applications* 96, 19 (2014), 1–8.

B. J. Sowmya, Chetan, and K. G. Srinivasa. 2016. Large scale multi-label text classification of a hierarchical dataset using Rocchio algorithm. In *Proceedings of the 2016 International Conference on Computation System and Information Technology for Sustainable Solutions*. IEEE, 291–296.

M. Srikanth, J. Varner, M. Bowden, and D. I. Moldovan. 2005. Exploiting ontologies for automatic image annotation. In *Proceedings of the SIGIR*. Ricardo A. Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat, and John Tait (Eds.), ACM, 552–558. Retrieved from http://dblp.uni-trier.de/db/conf/sigir/sigir2005.html#SrikanthVBM05.

Divya Srivastava, Rajitha Bakthula, and Suneeta Agarwal. 2019. Image classification using SURF and bag of LBP features constructed by clustering with fixed centers. *Multimedia Tools and Applications* 78, 11 (2019), 14129–14153.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

D. Storcheus, A. Rostamizadeh, and S. Kumar. 2015. A survey of modern questions and challenges in feature extraction. In *Proceedings of the Feature Extraction: Modern Questions and Challenges*. PMLR, 1–18.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: An affective extension of wordnet. In *Proceedings of the Lrec.* Vol. 4. Lisbon, 40.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL'19, Florence, Italy, July 28- August 2, 2019) Volume 1: Long Papers*, Association for Computational Linguistics, 3645–3650.

B. L. Sturm. 2012. A survey of evaluation in music genre recognition. In *Proceedings of the International Workshop on Adaptive Multimedia Retrieval*. Springer, 29–66.

T. Subramaniam, H. A. Jalab, and A. Y. Taqa. 2010. Overview of textual anti-spam filtering techniques. *International Journal of Physical Sciences* 5, 12 (2010), 1869–1882.

F. M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*. 697–706.

C. Sun, X. Qiu, Y. Xu, and X. Huang. 2019a. How to fine-tune bert for text classification?. In *Proceedings of the China National Conference on Chinese Computational Linguistics*. Springer, 194–206.

Y. Sun and S. Ghaffarzadegan. 2020. An ontology-aware framework for audio event classification. In *Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 321–325.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced representation through knowledge integration. CoRR abs/1904.09223 (2019).

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 1–9.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient Transformers: A Survey. *ACM Comput. Surv.* 55, 6 (2023), 109:1–109:28.

T. Theodorou, I. Mporas, and N. Fakotakis. 2014. An overview of automatic audio segmentation. *International Journal of Information Technology and Computer Science* 6, 11 (2014), 1.

Thirumoorthy Karpagalingam and Muneeswaran Karuppiah. 2021. Feature selection using hybrid poor and rich optimization algorithm for text classification. *Pattern Recognit. Lett.* 147 (2021), 63–70.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. 2022. ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training. *IEEE Transactions on Pattern Analysis Machine Intelligence* 01 (2022), 1–9.

Lloyd N. Trefethen and David Bau III. 1997. *Numerical Linear Algebra*. Vol. 50. Siam.

B. Trstenjak, S. Mikac, and D. Donko. 2014. KNN with TF-IDF based framework for text categorization. 69 (2014), 1356–1364. DOI: https://doi.org/10.1016/j.proeng.2014.03.129

P. D. Turney and P. Pantel. 2010. From frequency to meaning : Vector space models of semantics. 37 (2010), 141–188.

T. Tuytelaars and K. Mikolajczyk. 2008. Local invariant feature detectors: A survey. *Foundations and Trends® in Computer Graphics and Vision* 3, 3 (2008), 177–280.

J. Uys, N. Du Preez, and E. Uys. 2008. Leveraging unstructured information using topic modelling. In *Proceedings of the PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*. IEEE, 955–961.

Xavier Valero and Francesc Alias. 2012. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia* 14, 6 (2012), 1684–1689.

D. A. Van Dyk and X. Meng. 2001. The art of data augmentation. *Journal of Computational and Graphical Statistics* 10, 1 (2001), 1–50.

V. Vapnik and A. Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural Networks* 22, 5–6 (2009), 544–557.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 5998–6008.

S. Vijayarani, M. J. Ilamathi, and M. Nithya. 2015. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks* 5, 1 (2015), 7–16.

J. Vogel and B. Schiele. 2007. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision* 72, 2 (2007), 133–157.

D. Vrandečić and M. Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM* 57, 10 (2014), 78–85.

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 3360–3367.

Luyu Wang and Aaron van den Oord. 2021. Multi-format contrastive learning of audio representations. CoRRabs/2103.06508 (2021).

Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard F. Lyon, and Rif A. Saurous. 2017. Trainable frontend for robust and far-field keyword spotting. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5670–5674.

A. B. Warriner, V. Kuperman, and M. Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45, 4 (2013), 1191–1207.

C. Whittaker, B. Ryner, and M. Nazif. 2010. Large-scale automatic classification of phishing pages. In *Proceedings of the Network and Distributed System Security Symposium, (NDSS'10, San Diego, California, USA, 28th February - 3rd March 2010)*, The Internet Society.

L. Wu, S. C. Hoi, and N. Yu. 2010. Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing* 19, 7 (2010), 1908–1920.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 481–492.

Chun Yang, Chang Liu, and Xu-Cheng Yin. 2022. Weakly correlated knowledge integration for few-shot image classification. *Machine Intelligence Research* 19, 1 (2022), 24–37.

X. Yang, C. Macdonald, and I. Ounis. 2018. Using word embeddings in twitter election classification. *Information Retrieval Journal* 21, 2–3 (2018), 183–207.

Jingyi Ye, Xiaojun Jing, and Jia Li. 2017. Sentiment analysis using modified LDA. In *Proceedings of the International Conference on Signal and Information Processing, Networking and Computers*. Springer, 205–212.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. CoRR abs/1506.06579 (2015).

L. Younes, B. Romaniuk, and E. Bittar. 2012. A comprehensive and comparative survey of the SIFT algorithm-feature detection, description, and characterization. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, Vol. 2. SCITEPRESS, 467–474.

L. Yu, J. Wang, K. R. Lai, and X. Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 534–539.

Saadia Zahid, Fawad Hussain, Muhammad Rashid, Muhammad Haroon Yousaf, and Hafiz Adnan Habib. 2015. Optimized audio classification and segmentation algorithm by using ensemble methods. *Mathematical Problems in Engineering* 2015 (2015), 209814–209825.

N. M. Zaitoun and M. J. Aqel. 2015. Survey on image segmentation techniques. *Procedia Computer Science* 65 (2015), 797–806.

Masoumeh Zareapoor and K. R. Seeja. 2015. Feature extraction or feature selection for text classification: A case study on phishing email detection. *International Journal of Information Engineering and Electronic Business* 7, 2 (2015), 60.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. LEAF: A learnable frontend for audio classiffication. In *9th International Conference on Learning Representations (ICLR'21)*. Virtual Event, Austria, OpenReview.net.

M. D. Zeiler and R. Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*. Springer, 818–833.

D. Zhang, M. Cui, Y. Yang, P. Yang, C. Xie, D. Liu, B. Yu, and Z. Chen. 2019a. Knowledge graph-based image classification refinement. *IEEE Access* 7 (2019), 57678–57690.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *Proceedings of the Web Conference 2020*. 201–211.

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73, 2 (2007), 213–238.

Q. Zhang and S. Zhu. 2018. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.

S. Zhang, H. Tong, J. Xu, and R. Maciejewski. 2019. Graph convolutional networks: A comprehensive review. *Computational Social Networks* 6, 1 (2019), 1–23.

T. Zhang and C. J. Kuo. 1998. Hierarchical system for content-based audio classification and retrieval. In *Proceedings of the Multimedia Storage and Archiving Systems III*, Vol. 3527. International Society for Optics and Photonics, 398–409.

W. Zhang, T. Yoshida, and X. Tang. 2011. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications* 38, 3 (2011), 2758–2765.

Xinwei Zhang and Bin Wu. 2015. Short text classification based on feature extension using the n-gram model. In *Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE, 710–716.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL'19)* 1 (2019), 1441–1451.

Zhiling Zhang, Zelin Zhou, Haifeng Tang, Guangwei Li, Mengyue Wu, and Kenny Q. Zhu. 2021. Enriching ontology with temporal commonsense for low-resource audio tagging. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. 3652–3656.

Arman Zharmagambetov, Qingming Tang, Chieh-Chi Kao, Qin Zhang, Ming Sun, Viktor Rozgic, Jasha Droppo, and Chao Wang. 2022. Improved representation learning for acoustic event classification using tree-structured ontology. In *Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 321–325.

A. Zheng and A. Casari. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. "O'Reilly Media, Inc."

Yaguang Zhu, Chaoyu Jia, Chao Ma, and Qiong Liu. 2019. SURF-BRISK–based image infilling method for terrain classification of a legged robot. *Applied Sciences* 9, 9 (2019), 1779.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.

J. Zou, W. Li, C. Chen, and Q. Du. 2016. Scene classification using local and global features with collaborative representation fusion. *Information Sciences* 348 (2016), 209–226.

Received 18 May 2021; revised 24 December 2022; accepted 23 January 2023