Multimodal Technologies and Interaction

MDPI

Review

Harnessing AI and NLP Tools for Innovating Brand Name Generation and Evaluation: A Comprehensive Review

Marco Lemos [1,†], Pedro J. S. Cardoso [2,*,†] and João M. F. Rodrigues [2,†]

[1] Instituto Superior de Engenharia, Universidade do Algarve, 8005-139 Faro, Portugal; a72178@ualg.pt
[2] NOVA LINCS & Instituto Superior de Engenharia, Universidade do Algarve, 8005-139 Faro, Portugal; pcardoso@ualg.pt (P.J.S.C.); jrodrig@ualg.pt (J.M.F.R.)
[*] Correspondence: pcardoso@ualg.pt
[†] These authors contributed equally to this work.

Abstract: The traditional approach of single-word brand names faces constraints due to trademarks, prompting a shift towards fusing two or more words to craft unique and memorable brands, exemplified by brands such as SalesForce© or SnapChat©. Furthermore, brands such as Kodak©, Xerox©, Google©, Häagen-Dazs©, and Twitter© have become everyday names although they are not real words, underscoring the importance of brandability in the naming process. However, manual evaluation of the vast number of possible combinations poses challenges. Artificial intelligence (AI), particularly natural language processing (NLP), is emerging as a promising solution to address this complexity. Existing online brand name generators often lack the sophistication to comprehensively analyze meaning, sentiment, and semantics, creating an opportunity for AI-driven models to fill this void. In this context, the present document reviews AI, NLP, and text-to-speech tools that might be useful in innovating the brand name generation and evaluation process. A systematic search on Google Scholar, IEEE Xplore, and ScienceDirect was conducted to identify works that could assist in generating and evaluating brand names. This review explores techniques and datasets used to train AI models as well as strategies for leveraging objective data to validate the brandability of generated names. Emotional and semantic aspects of brand names, which are often overlooked in traditional approaches, are discussed as well. A list with more than 75 pivotal datasets is presented. As a result, this review provides an understanding of the potential applications of AI, NLP, and affective computing in brand name generation and evaluation, offering valuable insights for entrepreneurs and researchers alike.

Keywords: brandability; brand name generation; brand name evaluation; artificial intelligence; natural language processing

check for updates

Citation: Lemos, M.; Cardoso, P.J.S.; Rodrigues, J.M.F. Harnessing AI and NLP Tools for Innovating Brand Name Generation and Evaluation: A Comprehensive Review. *Multimodal Technol. Interact.* **2024**, *8*, 56. https://doi.org/10.3390/mti8070056

Academic Editor: Stephan Schlögl

Received: 23 May 2024
Revised: 19 June 2024
Accepted: 24 June 2024
Published: 1 July 2024

1. Introduction

In an era marked by rising entrepreneurship and the incessant quest for innovation, the process of brand name generation stands as a pivotal endeavor. The significance of crafting unique and memorable brand names cannot be overstated when considering their role as the initial touchpoint between businesses and consumers. However, with a plethora of startups emerging annually, the challenge of finding distinctive and brandable names has intensified. Traditional approaches for brand name generation, which often revolve around single-word names, face constraints due to trademark issues and often fall short in providing novel solutions, which opens doors for the exploration of innovative methodologies [1–3].

This review paper delves into the domain of artificial intelligence (AI)-enabled brand name generation and evaluation, leveraging the power of natural language processing (NLP) and affective computing (AffC) to assess the emotional impact of brand names. The fusion of AI, NLP, and AffC technologies and tools offers a promising path for creating unique and brandable names that echo within target audiences while adhering to legal and commercial considerations. In this context, the current state of the research field underscores the potential of AI, NLP, and AffC in reshaping brand naming practices. While existing brand name generators offer some utility, they often have proprietary algorithms and (seem to) lack the sophistication to analyze semantic nuances and sentiments effectively. This gap presents an opportunity for pioneering research aimed at developing AI models capable of discerning the brandability of multi-word combinations, neologisms, and invented names, thereby expanding the pool of viable brand names for emerging businesses. Key publications in the field have explored various aspects of brand name generation using NLP techniques. However, controversial hypotheses surrounding the subjective nature of brandability and the effectiveness of NLP algorithms warrant careful consideration while emphasizing the need for rigorous validation.

This study's main goal is to investigate instruments that can produce two-way AI models, with one capable of generating brand names and the other able to evaluate the brandability of the generated name, thereby facilitating the creation of unique and memorable brand names for businesses across diverse industries. By bridging the gap between linguistic analysis and commercial viability, our goal is to equip entrepreneurs with the essential tools required to create unique brand identities in a highly competitive market.

This research goes beyond the traditional approach of single-word brand names, which often face constraints due to trademarks, prompting a shift towards fusing two or more words to craft unique and memorable brands, exemplified by SalesForce© and SnapChat©. Furthermore, brands such as Ikea©, Kodak©, Xerox©, Google©, Häagen-Dazs©, and Twitter© have become household names although they are not real words, underscoring the importance of brandability in the naming process.

The main contribution of this paper is to provide a review of basic concepts, tools, and technologies that can be used to generate and evaluate brand names. We explore the datasets and techniques used to train AI models as well as the strategies for using objective data to validate the brandability of generated names. Special focus is given to the emotional and semantic aspects of brand names, which are often overlooked in traditional approaches. A pivotal list of datasets and AI models is presented as well.

In summary, we believe that this review will provide an initial comprehensive understanding of the potential applications of AI, NLP, and AffC in brand name generation and evaluation, offering valuable insights for entrepreneurs and researchers alike. This will allow future development of AI models capable of discerning the brandability of multi-word combinations, neologisms, and invented names, thereby expanding the pool of viable brand names for emerging businesses. Further, it can be used as a starting point for future research in the field, as it provides a wide range of pivotal concepts, tools, and technologies that can be used to generate and evaluate brand names.

The review starts by reviewing the literature on brand naming, focusing on the challenges and strategies involved in creating effective brand names (Section 2). Subsequently, it delves into the existing AI models and techniques for brand value extraction (Section 3.1) and text classification (Section 3.2), highlighting their potential applications in brand name generation and evaluation. Text generation and text-to-speech models are discussed in the context of brand name creation (Sections 3.3 and 3.4). Section 4 explores the datasets used to train AI models for brand name generation and evaluation. Finally, this document discusses the implications of AI in branding and outlines future research directions in the field.

2. Brand Naming

The naming of brands has been a subject of interest to researchers for many years, as it plays a critical role in consumer awareness and product image. Despite the significant progress made in branding literature, naming a brand remains a challenging task, with debates persisting on whether there has been significant progress in the process. The recent literature has extensively discussed trademark-related issues involved in brand naming.

The need for unique and memorable brand names has led to the emergence of new naming strategies, such as the use of neologisms, invented names, and multi-word combinations. The importance of domain names in the digital age has been highlighted as well, with the value of a domain name being determined by its ability to attract visitors and convert them into paying customers. For example, the expansion of the internet has brought about a growing demand for web names, with cybersquatting becoming a problem, further emphasizing the importance of choosing the right domain name for a business.

In addition to highlighting the potential challenges that entrepreneurs may encounter during the naming process, Eskiev et al. [3] stressed the importance of naming for establishing a favorable image of a company within the market. They examined a range of naming strategies and tactics, including personal branding, neologism-based naming, shortened naming, reference naming, associative naming, and more. They also addressed the benefits and drawbacks of various name strategies, offering instances of well-known brands that have employed particular naming conventions.

Arora et al. [1] suggested a framework of brand name classification, stating that brief, easy to pronounce, and memorable names that are able to communicate product positioning are necessary for effective branding. In addition, brand names ought to be memorable, likeable, and legally protected. Thus, generally speaking, naming categories are often made up of amalgamations that are unexpected (names using common words in unexpected ways, e.g., Apple$^{©}$), emotive (Crunchie$^{©}$), creative (purely novel names invented by organizations, e.g., Exxon$^{©}$), or semantically-based (names that convey information about the product, e.g., Burger King$^{©}$). The authors investigated frameworks such as McCune's classification and the Juliet and Joyce principles, emphasizing that the Joyce principle emphasizes the use of sound symbolism in communicating meaning, the Juliet principle contends that a name's spoken form is meaningless, and McCune's classification includes semantically-based names, person names, unexpected names, and invented names.

According to Moro-Visconti [2], a domain's value is based on how well it attracts visitors and turns them into paying customers, which is a complex process that is dependent on a number of factors. As a matter of fact, the intrinsic value of a domain name is a small but essential part of the entire web value chain, which includes a portfolio of web intangibles. However, Moro-Visconti asserts that the characteristics of marketable domain names can be summed up as follows: (a) good top-level domain (preferably ".com", the most popular and easily remembered); (b) short length (shorter domains are more valuable due to their rarity); (c) radio test (easily understood and spelled after hearing it); (d) correct spelling (proper spelling is essential for value); (e) meaningful keywords (these should be in-demand and related to the website's content); (f) easy to remember; (g) brandable (having a nice pronunciation, interesting combination of letters, or appealing visual effect); (h) no hyphens or numbers; and (i) popular keywords.

In summary, Eskiev et al., Arora et al., and Moro-Visconti have contributed distinct perspectives on naming and branding. While Eskiev et al. and Arora et al. highlight the significance of a brand's name in shaping its image and offer strategies and frameworks for effective naming, Moro-Visconti's focus lies on the value of domain names and their role as virtual "shop windows". Each paper provides valuable insights into various aspects of naming and branding, collectively offering a comprehensive understanding of the importance of a brand's name and domain name in influencing its image and market perception. The next sections delve into AI models and techniques that can be used to generate and evaluate brand names while leveraging the power of NLP and AffC to assess the names' emotional impact. The fusion of AI, NLP, and AffC technologies and tools offers a promising avenue for creating unique and brandable names that resonate with target audiences while adhering to legal and commercial considerations.

3. AI Models and Techniques

Our objective in this part is to investigate two types AI models, specifically, methodologies capable of (a) generating brand names and (b) evaluating the brandability of the proposed or generated names. As no single article has yet carried out these precise tasks, we chose to search for articles that could be of assistance in adhering to our aims. Research was conducted by looking for articles on Google Scholar, IEEE Xplore, and ScienceDirect using the following keywords: "brand name generation", "brand name evaluation", "brand name assessment", "brand name creation", "brand name AI", "brand name NLP", "brand name sentiment analysis", "brand name classification", "brand name value", and "brand name semantic". The same queries were made without the word "brand" to ensure that no relevant articles were missed. Then, modernness was added as a criterion for the selection of articles, with older articles included only if they were considered relevant or seminal or if no alternatives were found. The top results from each search engine were considered. Due to the broadness of the search, we searched for articles on semantic assessment, naming, text production, and brand value extraction for generating business names as well as on voice classification, pronunciation, sentiment assessment, and text classification for assessing company names.

The remaining section is divided into six parts: brand value identification; text classification; text generation; text-to-speech; and some concluding critical remarks. A summary arranged by application is provided in Table 1.

Table 1. Summary of discussed works by category.

Brand value extraction	[4,5]
Text sentiment analysis	[6–11]
Semantic analysis	[12–16]
Text generation	[6,17–21]
Text-to-speech	[22,23]
Speech classification	[24–26]
Speech sentiment analysis	[27–29]

3.1. Brand Value Identification

In e-commerce systems, it is vital to extract information about product attribute values. The brand is one of the most important aspects of a product, as it often a large influence on consumers' behavior and purchasing decisions. Therefore, in order to address the primary difficulty of finding new brand names, it is imperative that brand information be accurately extracted and assessed.

Sabeh et al. [4] presented OpenBrand, a model architecture designed for extracting brand values from unstructured product titles. By incorporating both word and character-level embeddings, OpenBrand outperforms existing state-of-the-art models across various categories. The model demonstrates strong generalization capabilities, particularly in zero-shot extraction scenarios, and performs well even with compound brand values. Further, the problem of unseen words in brand names, which can arise due to sub-branding, brand fragmentation, or emerging businesses, is addressed. Additionally, the authors provided a large real-world dataset [30] derived from [31] specifically focusing on brand names. The strategy suggested by Sabeh et al. for brand value identification employs a generative language model to identify brand values from product descriptions [5]. Unlike previous methods that rely on sequence tagging and fail to identify brand values that are not explicitly mentioned, the proposed method formulates the task as a sequence-to-sequence (S2S) generation problem. The author fine-tuned a generative language model, specifically T5 [32], by incorporating the product category into the input. This category-aware model, named GAVI, was trained on a dataset containing over 250,000 product titles across various categories [30], and according to the author was able to routinely beat rival baselines in all product categories.

3.2. Text Classification

The process of grouping texts into structured categories based on their content and context is known as text classification (TC). Sentiment analysis, news classification, semantic analysis, and subject classification are examples of common TC tasks. By employing advanced machine learning (ML) techniques such as deep learning (DL), researchers have recently demonstrated that it is effective to cast many natural language understanding tasks as TC tasks, including question answering and natural language inference [33–35]. A selection of the stated TC tasks are covered in the sections below.

3.2.1. Text Sentiment Analysis

Text sentiment analysis implies analyzing textual data, such as tweets, movie reviews, or product reviews, to ascertain the polarity and perspective of people's opinions. This task can be framed as either a binary (or trinary) classification problem, where texts are classified into positive or negative (or neutral) categories, or as a multi-class problem in which finer-grained labels or intensities are assigned to the expressed sentiment (e.g., happy, sad, angry).

Nurmambetov et al. [6] proposed a logistic regression model for binary sentiment analysis, starting with language identification and tokenization using tools from the Natural Language Toolkit (NLTK) [36]. They utilized NLTK's stop-word corpus to eliminate non-informative words after tokenization, followed by lemmatization with the StanfordNLP library [37]. CountVectorizer was then applied to transform textual data into a numerical format, considering both term frequency (TF) and inverse document frequency (IDF).

Muhammad et al. [7] described a sentiment analysis model based on Word2vec supported on the word similarity in the corpus and adjacency to one another. It creates a vector space, called the Word2vec space, after which a Long Short-Term Memory (LSTM) network is introduced. The model was applied to Indonesian hotel reviews. The authors tested various Word2vec and LSTM setups and found that the continuous skip-gram architecture provided higher accuracy compared to continuous bag-of-words (CBOW) in the Word2vec design. On the other hand, according to Mikolov et al., the order of words in the history had no bearing on the projection in the CBOW architecture [38]. Despite this, Mikolov et al. claimed that architectures comparable to CBOW attempt to maximize word categorization based on a word in the same phrase rather than guessing the current word based on context. Notably, the dataset that Mikolov et al. examined was derived from the collection of hotel reviews available on the Traveloka website (https://www.traveloka.com, accessed on 26 April 2024).

In this context, lemmatization (e.g., [6]) or stemming (e.g., [7]) are two options for preprocessing. Words are normalized by using one of these techniques to make them uniform across formats. Both are defined by Khyani et al. as processes that create variants of a root/base word [39]. On one hand, stemming techniques are applied by cutting off the beginning or end of the word-text and recording frequently used prefixes and suffixes that may be present in an inflected word-text. On the other hand, lemmatization takes into consideration the analysis of the word-text with the goal of finding something that provides it additional meaning. Even though lemmatization is a more difficult process than stemming, Khyani et al. contend that lemmatization is ultimately the better option [40].

Jiang et al. [8] introduced a hybrid Word2vec–LSTM model to assess the sentiment of movie reviews. The Word2vec network layer was employed to extract features from the text, mining the word context semantics to obtain the text vector, which was then used as the input of the LSTM layer network. They also created two types of benchmark models by combining text vectorization with various ML techniques such as, random forest (RF), logistic regression (LR), support vector machine (SVM), *k*-nearest neighbors (KNN), and hash trick (HT). Their study found that the Word2vec–LSTM hybrid model outperformed the benchmark models in terms of prediction results. The binary sentiment analysis dataset utilized in this work was the movie review collection presented in [41], which primarily consists of 25,000 positive and 25,000 negative film review texts.

Sentiment analysis of the transgender community from social media data was the main topic of a study by Liu et al. [9]. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) deep learning models were compared with ML classifiers such as KNN, LR, RF, SVM, and naive Bayes. In this case, the DL models utilizing Word2vec embeddings outperformed the other classifiers.

In order to address the multi-class review rating classification problem, Hassan et al., [10] presented the Domain-Specific Word Embeddings–Gated Recurrent Unit (DSWE–GRNN) architecture. Due to the rarity of reviews from the same person for estimating the strictness towards a particular sentiment, Hassan et al. used domain-specific word embeddings, an approach that is independent of the reviewer's information. An RNN-based architecture efficiently and effectively trains the model for review rating classification by extracting the hidden contextual information from the domain-specific word embeddings. They assessed their model using the Hotel Reviews [42] and IMDB [43] datasets.

Valence-arousal dimensional space is a common way of representing emotions on a continuous scale, offering a nuanced way to distinguish between different emotional states. In this context, Mendes et al. [11] delved into the realm of emotion analysis by focusing on the quantification of valence and arousal in textual data using multilingual pretrained transformers. Their study emphasized the significance of model size in accurately predicting valence and arousal, showcasing the impact of fine-tuning large models for this task. By compiling 34 publicly available psycho-linguistic datasets from various languages into a unified dataset (see Table 2), their research evaluated the performance of multilingual DistilBERT [44] and XLM-RoBERTa [45] models in predicting affective ratings from textual content. The assessment took into account how model size affects the accuracy of predictions; in order to determine how model complexity influences the precision of valence and arousal predictions, several transformer model sizes were tested. Experiments excluding specific datasets, such as COMETA Stories [46] and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [47], led to performance improvements in valence and arousal predictions across different languages, indicating the influence of dataset composition on model performance. The study also evaluated the generalization capability of the best model to languages not included in the training data, such as Polish and Portuguese, through zero-shot evaluations. The models showed promising results in predicting valence and arousal for these languages as well, despite not being extensively trained on them.

The papers presented in this section show diverse sentiment analysis techniques across various domains and languages, employing models such as LR, Word2vec–LSTM hybrids, and DL architectures and achieving state-of-the-art results in sentiment classification tasks. Despite the varied methodologies and datasets used, each study highlights the need for tailored approaches to sentiment analysis, emphasizing the absence of a universal solution due to factors such as differing domain complexities.

Table 2. List of datasets used in the discussed works and the tasks they were applied to. Non-English datasets are identified.

Dataset	Description	Applied to
AG News [48]	The academic news search engine ComeToMyHead gathered news stories from over 2000 news sources to create the AG News dataset. Samples were brief sentences and used four-class classification.	News classification
Amazon [49]	Labels for multi-class (five) classification as well as binary classification. The Amazon binary classification dataset has 3,600,000 reviews for training and 400,000 for testing.	Sentiment analysis
ANEW-It [50]	Affective norms for words rated in terms of valence, arousal, and dominance (in Italian).	Valence and arousal computation (e.g., in [11])
ANEW-Pt [51]	Valence, arousal, and dominance ratings for 1034 words (in Portuguese).	Valence and arousal computation (e.g., in [11])
ANGST [52]	Total of 1003 words (in German) evaluated based on imageability, potency, dominance, and arousal.	Valence and arousal computation (e.g., in [11])
ANPST [53]	Affective ratings for valence, arousal, dominance, origin, subjective significance, and source dimensions for 718 short texts (in Polish).	Valence and arousal computation (e.g., in [11])
ANPW_R [54]	Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability, and age of acquisition for 4900 words (in Polish).	Valence and arousal computation (e.g., in [11])
BAWL-R [55]	Includes evaluations for emotional valence, imageability, and emotional arousal (in German).	Valence and arousal computation (e.g., in [11])
Binary animacy collection [56]	Animacy ratings for all words in a word pool.	Assessment of word memorability (e.g., in [15])
COMETA [46]	A database of conceptual metaphors incorporating emotional ratings alongside linguistic properties, with stimuli including natural stories and isolated sentences, assessed for attributes such as valence, arousal, and metaphoricity (in German).	Valence and arousal computation (e.g., in [11])
Concreteness ratings [56,57]	Concreteness ratings on a five-point scale.	Assessment of word memorability (e.g., in [15])
Ćoso et al. [58]	Ratings of valence, arousal, and concreteness for 3022 words (in Croatian).	Valence and arousal computation (e.g., in [11])
CVAI [59]	Extended NTU irony corpus, which includes valence, arousal, and irony intensities on the sentence level and valence and arousal intensities on the context level (in Mandarin).	Valence and arousal computation (e.g., in [11])
CVAT [60]	Total of 2009 phrases taken from online sources rated for valence and arousal (in Mandarin).	Valence and arousal computation (e.g., in [11])
CVAW [60,61]	A sentimental vocabulary of 1653 terms with scores for valence and arousal. (in Mandarin).	Valence and arousal computation (e.g., in [11])
DBpedia [62]	Multilingual knowledge base assembled from Wikipedia's most popular infoboxes. Each sample in the most widely used version of DBpedia has a 14-class label.	News classification
Eilola et al. [63]	Valence, emotional charge, offensiveness, concreteness, and familiarity ratings for 210 nouns (in Finnish).	Valence and arousal computation (e.g., in [11])

Table 2. *Cont.*

Dataset	Description	Applied to
EmonBank [64,65]	Text corpus manually annotated with emotions according to the psychological valence–arousal–dominance scheme (in English).	Valence and arousal computation (e.g., in [11])
EmoTales [66]	Annotated corpus oriented to the narrative domain which uses two different approaches to represent emotional states, emotional categories, and emotional dimensions (in English).	Valence and arousal computation (e.g., in [11])
EUR-Lex [67]	Several categories of documents are included in the EUR-Lex dataset. This dataset's most-used version has 3956 categories and 19,314 documents based on various parts of EU law.	Topic classification
Facebook Posts [68]	Social media posts rated by two psychologically trained annotators on two separate ordinal nine-point scales of valence and arousal.	Valence and arousal computation (e.g., in [11])
FAN [69]	Affective norms for 1031 words rated on emotional valence and arousal (in French).	Valence and arousal computation (e.g., in [11])
FEEL [70]	Valence, arousal, and imagery ratings for 835 words (in French).	Valence and arousal computation (e.g., in [11])
Fisher [71]	Corpus containing conversational telephone speech of more than 16,000 English conversations.	Speech sentiment analysis (e.g., [29])
Hotel Reviews [42]	Dataset comprising 14,895 data samples (reviews) divided into five classes.	Multi-class sentiment classification (e.g., in [10])
IEMOCAP [47]	Total of 151 videos of recorded dialogues annotated for the presence of nine emotions (angry, excited, fearful, sad, surprised, frustrated, happy, disappointed, neutral) as well as valence, arousal, and dominance.	Speech sentiment analysis (e.g., [27])
IMDB [72]	Dataset consisting of 50,000 data samples (movie reviews) divided into ten classes	Multi-class sentiment classification (e.g., in [10])
Irony [73]	Composed of the arXiv collection, Twitter dataset for topic classification of tweets, and annotated comments from the social news website Reddit.	Topic classification
Kapucu [74]	Norms of valence and arousal for 2031 words (in Turkish).	Valence and arousal computation (e.g., in [11])
LANG [75]	Total of 1000 nouns rated for emotional valence, arousal, and concreteness (in German).	Valence and arousal computation (e.g., in [11])
Librilight [76]	Corpus consisting of 60,000 h of English speech with over 7000 unique speakers.	Text-to-speech (e.g., [77])
Librispeech [78]	LibriSpeech is a corpus of approximately 1000 h of read English speech.	Text-to-speech (e.g., [22])
LibriTTS [79]	Multi-speaker corpus of approximately 585 h of read English speech at a sampling rate of 24 kHz.	Text-to-speech (e.g., [77])
MAS [80]	Dimensional and categorical measures of emotion were used to score 192 sentences (in Portuguese).	Valence and arousal computation (e.g., in [11])
Moors et al. [81]	Norms of valence, arousal, dominance, and age of acquisition for 4300 words (in Dutch).	Valence and arousal computation (e.g., in [11])

Table 2. *Cont.*

Dataset	Description	Applied to
Movie Review [82]	A collection of movie reviews created with the intention of identifying the sentiment attached to each review; 10,662 sentences are included, with an equal number of positive and negative samples.	Sentiment analysis
MPQA [83]	Opinion corpus classified into two classes; 10,606 sentences were taken from news items pertaining to a broad range of news sources.	Sentiment analysis
MS MARCO [84]	Set of questions sampled from user searches and passages from actual web documents. Includes generative replies.	Question answering
MSRP [85]	Total of 1725 samples for testing and 4076 samples for training in MSRP. Every example consists of two statements either labeled as either paraphrases or lacking a binary label.	Natural Language Inference (NLI)
Multi-NLI [86]	Total of 43,300 sentence pairings, each with appended textual entailment labels. The corpus is an expansion of SNLI spanning a larger variety of spoken and written text genres.	Natural Language Inference (NLI)
NAWL [87]	A database with 2902 Polish words, including nouns, verbs, and adjectives along with ratings of emotional valence, arousal, and imageability (in Polish).	Valence and arousal computation (e.g., in [11])
NRC-VAD [88]	Valence, arousal, and dominance ratings for 20,000 English words (in English).	Valence and arousal computation (e.g., in [11])
Ohsumed [89]	Total of 7400 documents, each consisting of a medical abstract labeled by one or more classes chosen from 23 categories related to cardiovascular disorders.	Topic classification
One-Billion word [90]	Dataset consisting of approximately one billion words extracted from various sources on the internet, such as news articles, blogs, and websites.	Text generation (e.g., [20])
Open-Brand [30]	Dataset containing over 250,000 product brand–value annotations, with more than 50,000 unique values across eight main categories of Amazon product profiles.	Brand values extracted from product descriptions (e.g., in [4,5])
PANIG [91]	Psycholinguistic and affective standards for 619 colloquial terms. Valence, arousal, familiarity, semantic transparency, figurativeness, and concreteness were evaluated for each phrase (in German).	Valence and arousal computation (e.g., in [11])
PubMed [92]	Collection of documents from medical and biological research publications. Every document has been tagged using the MeSH set classes. A sentence's function in an abstract is indicated by labeling it with one of the following classes: background, objective, method, outcome, or conclusion.	News classification
Quora [93]	Dataset consisting of over 400,000 lines of potential question duplicate pairs.	Text generation (e.g., [19])
Recall Memory [56,94]	A total of 98 participants attended 23 experimental sessions. In each session, participants looked over 24 lists, each of 24 words. Participants received 24 s to respond to basic math problems before having 75 s to memorize as many words as they could from the just-presented list.	Assessment of word memorability and estimation of word recall (e.g., in [15])

Table 2. *Cont.*

Dataset	Description	Applied to
Recognition Memory [95,96]	A total of 171 subjects participated in up to 20 experimental sessions each. Participants examined 12–16 lists of 16 items. Participants finished a recognition memory task at the conclusion of each session by marking whether each of the words had been shown earlier.	Assessment of word memorability and estimation of word recognition (e.g., in [15])
Reuters news [97]	One of the most widely used datasets for text classification, gathered in 1987 from the Reuters financial newswire service.	News classification
Speechocean762 [98]	Open-source speech corpus designed for pronunciation assessment use, consisting of 5000 English utterances from 250 non-native speakers. Five experts annotated each of the utterances at the sentence level, word level, and phoneme level.	Assessment of pronunciation (e.g., [26])
S.-Gonzalez et al. [99]	Norms of valence and arousal for 14,031 words (in Spanish).	Valence and arousal computation (e.g., in [11])
SICK [100]	10,000 pairs of English sentences that have been labeled as neutral, entailment, or contradiction.	Natural Language Inference (NLI)
SNLI [101]	A dataset of 570 k English sentence pairs labeled for entailment, contradiction, and neutrality, serving as a benchmark for text representation evaluation and NLP model development.	Natural Language Inference (NLI)
Söderholm [79]	Valence and arousal ratings for 420 Finnish nouns by age and gender (in Finnish).	Valence and arousal computation (e.g., in [11])
Sogou News [102]	The SogouCA and SogouCS news corpora are combined to create the Sogou News dataset. The news items' domain names in the URL define their classification labels.	News classifications
SST [103]	There are two versions available, called SST-1 and SST-2, respectively, one with binary labels and the other with fine-grained labels (five-class). 11,855 movie reviews make up SST-1. SST-2 is divided into three training, development, and test sets, each with a size of 6920, 872, and 1821, respectively.	Sentiment analysis
SUBTLEX-US [104]	A dataset that shows the proportion of movies that contain a given word.	Assess word memorability (e.g., in [15])
SWBD-sentiment [105]	This corpus contains a total of 49,500 labeled utterances covering 140 h of audio. Each sentiment label in this corpus can be one of three options: positive, negative, and neutral.	Speech sentiment analysis (e.g., [27,29])
The Glasgow Norms [106]	A set of normative ratings for 5553 words on 9 psycholinguistic dimensions: arousal, valence, dominance, concreteness, imageability, familiarity, age, semantic size, and gender association.	Valence and arousal computation (e.g., in [11])
Top 500 companies by net impact [107]	This dataset includes all companies on the Fortune Global 500 list, 2020 edition, i.e., 500 largest companies in the world by revenue. All companies have been ranked based on their net impact.	
Top 100 Global Brands [108]	Top 100 companies in 2022, including 2022 and 2021 rating, company name, company founder, year of establishment, industry, location, website, additional key people, and 2022 and 2021 brand value.	

Table 2. *Cont.*

Dataset	Description	Applied to
TREC-QA [109]	Two versions: TREC-6 and TREC-50 with 6 and 50 classes, respectively. Both have 5452 training examples and 500 test examples.	Question answering
TTS-Portuguese [110]	The dataset has approximately 10 h and 28 min of speech from a single speaker, recorded at 48 Khz, containing a total of 3632 audio files in Wave format (in Portuguese).	Text-to-speech (e.g., [77])
USF free association norms [111]	Word pool consisting of 576 words.	Assess word memorability (e.g., in [15])
VCTK [112]	Corpus including speech data uttered by 110 English speakers with various accents. Each speaker reads out about 400 sentences, which were selected from a newspaper.	Text-to-speech (e.g., [77])
Verheyen et al. [113]	Norms for 1000 Dutch adjectives, covering lexicosemantic variables such as age of acquisition, familiarity, concreteness, and imageability alongside affective variables such as valence, arousal, and dominance as well as distributional variables (in Dutch).	Valence and arousal computation (e.g., in [11])
Warriner [114]	Valence, arousal, and dominance ratings on a nine-point scale.	Assessment of word memorability (e.g., in [15])
WOS [115]	Data and metadata of published papers.	Topic classification
XANEW [114]	Norms of valence, arousal, and dominance for 13,915 English lemmas.	Valence and arousal computation (e.g., in [11])
Xu [116]	Valence and arousal ratings for 11,310 simplified words (in Mandarin).	Valence and arousal computation (e.g., in [11])
Yee [117]	Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns (in Cantonese).	Valence and arousal computation (e.g., in [11])
Yelp [118]	Subset of businesses, reviews, and user data.	Text generation (e.g., [20])
7+ Million Company [119]	Dataset of over seven million companies, including company name, Linkedin URL, domain and industry, company size from 1–10,000+, company location, number of employees, and year of establishment.	
20 newsgroups [120]	Approximately 20,000 newsgroup documents partitioned (nearly) evenly across 20 different newsgroups.	Semantic analysis (e.g., in [121])
22.9+ Million Company [122]	Dataset of over 22.9+ million companies, including company name, Linkedin URL, domain and industry, company location, number of employees, and year of establishment.	

3.2.2. Semantic Analysis

Semantic analysis is an essential aspect of NLP techniques, effectively delivering the context of a sentence or a paragraph. For example, Wang et al. [12] presented a comprehensive approach for implementing an algorithm that checks for duplicated text based on semantic natural language analysis. Their models integrated advanced technologies such as Word2vec and latent Dirichlet allocation (LDA). The algorithm flow proposed by the authors is divided into two main parts, namely, training the topic model and searching for duplicate documents. Each time a new text is input, the model is updated and topic words extracted from the model are used to calculate the text similarity, ensuring the algorithm's efficiency. Cosine similarity is utilized for calculating the repetition rate in semantic text duplicate analysis, providing a robust measure of similarity between documents.

Maksutov et al. [13] delved into the scope of NLP and its role in comprehending quasi-structured or unstructured data, ultimately integrating it into a knowledge base utilizing a graph database. The authors underscored the challenges inherent in developing a universal solution for NLP, primarily due to the diverse grammatical, syntactic, and semantic forms across languages. They also delineated two fundamental approaches to tackling NLP tasks: rule-based algorithms and ML-based algorithms. Rule-based algorithms operate on grammatical rules and heuristics, while ML algorithms embrace training data to extract structures from text.

A method for natural language semantic understanding was developed by Li et al. [14]. The author outlined advancements in NLP with a focus on pretrained models, word sense disambiguation, and semantic integrity analysis. The work begins by highlighting NLP's importance in comprehending human language using computers, especially in the context of the internet, artificial intelligence, and big data. The advent of deep learning, particularly models such as BERT [123] and XLNet [124], has significantly improved various NLP tasks such as machine translation and sentiment analysis. The article discusses improved models based on BERT, including enhancements in generating tasks, knowledge integration, multi-task learning, masking modes, and training methods. XLNet is introduced as a model that addresses bidirectional context information using mechanisms such as permutation language modeling, two-stream self-attention, and loop mechanisms, showing significant improvements over previous models.

Aka et al. [15] studied the semantic determinants of memorability, investigating why certain words are more memorable than others. They used predictive ML models applied to word recognition and recall datasets, namely, recognition memory [95,96], recall memory [56,94], and wordpool [111]. They also used additional memory tasks to test generalization [125], human predictions, concreteness [56,57], valence and arousal [114], word frequency [104], and binary animacy [56]. Semantic features were extracted from the datasets to represent the underlying meaning of each word, including word frequency, animacy, valence, concreteness, arousal, and word length as potential predictors of memorability. The model was trained using ML algorithms on recognition and recall datasets, incorporating various algorithms such as Lasso regression, SVM, and RF.

Tuckute et al. [16,126] delved into the factors influencing word memorability, particularly focusing on the relationship between words and meanings. Their study conducted two large-scale experiments, each with over 1000 participants, to explore the memorability of words based on synonymy and homonymy. The experiments involved repeat detection tasks where participants identified repeated words in a series. The authors employed 2222 terms from the Subtlex corpus [104] and a new set of hand-picked terms. The results indicated that specific semantic categories, such as famous landmarks and games, were more memorable than others, such as weather and building components. Furthermore, when additional word norms were included in the model, such as valence, imageability, familiarity, concreteness, and arousal, the correlation increased, almost reaching the theoretical maximum of 0.65. These results indicate that a model trained on linguistic factors such as synonymy and homonymy can effectively predict word memorability across different sets of words and semantic categories.

Maulud et al. [127] provided an overview of semantic analysis within NLP, emphasizing its importance in interpreting the context of sentences and paragraphs. They explored the role of sentiment analysis and opinion mining, which are crucial for understanding user emotions and opinions expressed in text. Their paper discussed various NLP techniques, presenting an overview list of the most recent NLP techniques.

As a summary, the works addressed in this section concentrate on various facets of language understanding and natural language processing. While they are united by the shared goal of determining significant insights from textual data, their particular approaches and uses are different. To prepare textual material for analysis, all of the discussed works use preprocessing techniques such as tokenization, word embedding, and semantic analysis. Additionally, they all make use of ML or DL models for a variety of tasks, including topic modeling, text categorization, and word memorability prediction. However, the focus and outcomes differ, ranging across topic modeling, text similarity computation, text classification using semantic features, question answering systems, and assessing the memorability of words using semantic representations.

3.3. Text Generation

Text generation in AI models refers to the process of creating new textual content based on an existing dataset or prompt. It is a type of NLP task that incorporates ML algorithms, particularly deep learning, to generate coherent and contextually relevant text. Text generation can be used for various purposes, including: (a) language modeling—building a statistical model of language, which can be used for tasks such as machine translation or question-answering systems; (b) chatbots and virtual assistants—generating conversational responses for chatbots and virtual assistants that mimic human-like interactions; (c) text summarization—condensing longer documents or articles into shorter summaries while preserving the key information; (d) creative writing—generating creative and original content, such as stories, poems, or song lyrics; and (e) code generation—generating code snippets or entire programs based on a given specification or example. However, text generation also raises ethical concerns, such as the potential for generating biased or harmful content, as well as issues related to copyright and intellectual property [128].

Both supervised and unsupervised learning are possible in text generation models. Under supervised learning, a model is trained to predict a word or series of words based on an initial input (e.g., used in machine translation or picture captioning tasks). Unsupervised learning is the process by which a model learns to anticipate the following word in a sentence or paragraph by looking at the words that come before it [129,130] (e.g., when labeled data are unavailable, as is the case in language modeling jobs where large amounts of text data are accessible but there are no explicit labels).

Nurmambetov et al. [6] proposed a one-layer LSTM-based model for generating names. It utilizes preprocessing methods such as normalization and one-hot encoding with the Kazakh alphabet. The procedure consists of multiple phases and produces random (category, line) pairs, where the category is represented by a one-hot vector. Based on the current letter and the hidden state, the system guesses the next letter at each timestep.

Systems that can respond to queries in natural language, enhance the method of passage retrieval, improve the accuracy of natural language processing for factoid question-answer (QA) systems, and create a more workable system are among the research gaps noted in a study by Sadhuram et al. [17]. The proposed implementation phase involves using AI and NLP for better understanding and answering of questions. For example, their QA system includes question processing, passage retrieval, sentence retrieval, answer processing, and text summarization. The system was tested on over 422 articles of the SQUAD [131] dataset and approximately 87,599 cross-domain questions.

Song et al. [18] proposed the Emotional Dialogue System (EmoDS) method along with some ablation variants. EmoDS is a dialogue system designed to generate emotionally expressive responses in a coherent manner. The method is based on the S2S framework [132,133], which is extended with a lexicon-based attention mechanism to incorporate emotional words into the generated responses at the appropriate time steps. Additionally, a sequence-level emotion classifier is used to recognize emotional expressions even when no explicit emotional words are present. In short, the EmoDS architecture consists of a bidirectional LSTM-based (BiLSTM) encoder that encodes the input post into its vector representation, which is then used to initialize a decoder, which generates the response with a specific emotion with the assistance of an emotion classifier and lexicon-based attention mechanism. The BiLSTM classifier was trained on the NLPCC dataset [134].

Liu et al. [19] proposed a paraphrase generation model which is an extension of the traditional S2S model incorporating an attention mechanism and utilizing topic words as prior knowledge to improve the quality of generated paraphrases. In order to improve the semantic reference for rewriting, the model includes a way to add subject terms, which are extracted using LDA. The proposed method was trained using a combination of generative loss and direct supervision. The model was evaluated on two datasets, namely, Quora [93] and Twitter [135].

In the work by He et al. [20], lexically constrained text creation incorporates a few prespecified keywords into the output in an attempt to control the created text. According to the authors, their approach has high computing complexity and often produces generic or grammatically incorrect sentences. In this context, Constrained BART (CBART) for lexically constrained text creation is suggested as a solution to these problems. By splitting this effort into two smaller tasks, CBART divides the generation responsibility between the decoder and the encoder by utilizing a pretrained BART model, resulting in improved sentence quality. The outcomes of experiments conducted on One-Billion Words [90] and Yelp [118] databases demonstrate that CBART can produce believable text that is both diverse and high in quality while increasing the speed of inference. He et al. compared their proposed model with several strong baselines for lexically constrained text generation, including three traditional baselines (sep-B/F, asyn-B/F [136], and Grid Beam Search (GBS) [137]) and three more recent models (Constrained Sentence Generation via Metropolis–Hastings (CGMH) [138], POINTER [139], and XLNet-based Markov Chain Monte Carlo (X-MCMC) [140]).

Wang et al. [21] focused on S2S constrained text creation, in which the text generator is restricted to mentioning particular terms in the generated outputs that are inputs to the encoder. Although they can be trained to transfer surface tokens from encoders to decoders, pretrained S2S models such as T5 or transfer mechanisms cannot ensure constraint compliance. The authors proposed Mention Flags (MF), which track whether generated outputs of an S2S decoder satisfy lexical requirements. High constraint satisfaction is ensured, as MF models are trained to create tokens until all constraints are satisfied. In terms of dataset, this work used three different datasets for experiments: Commonsense Generative Reasoning (CommonGen) [141], End-to-End Natural Language Generation (E2E NLG) [142], and Novel Object Captioning at scale (nocaps) [143].

The works discussed in this section focus on text generation tasks. They share similarities in their objectives of generating coherent and contextually relevant text, but differ in their specific methodologies and applications. All papers employed DL models such as LSTM, S2S, or BART for the text generation tasks. They employed various preprocessing techniques to improve the quality of the generated text, such as one-hot encoding, attention mechanisms, and topic modeling. However, they differed in terms of focus and outcomes, with the former including name generation, enhancing passage retrieval, and question-answering and the latter involving the generation of emotionally expressive responses, paraphrases using topic words as prior knowledge, lexically constrained text, and S2S constrained text.

3.4. Text-to-Speech

Text-to-speech (TTS) refers to the process of converting written text into spoken words or audio. This technology enables computers, smartphones, and other devices to read text out loud. TTS systems often utilize NLP and ML techniques to generate human-like speech, which can be customized to sound like different genders, ages, or accents. TTS has a wide range of applications, including accessibility for people with visual impairments, audiobooks, navigation systems, virtual assistants, and more [34,144,145]. The quality of TTS systems is evaluated based on factors such as the naturalness, intelligibility, and expressiveness of the generated speech. From an affective computing perspective, it can serve as a signal to assess whether text is pleasing to human listeners, memorable and comprehensible, and evokes emotional responses. This fact is particularly relevant in the context of brand names, where the sound and pronunciation of a name can influence its brandability and memorability, as discussed in, e.g., [146,147].

The framework proposed by Latif et al. [22] is a system that generates synthetic emotional speech using a Tacotron-based TTS system. The system enhances speech emotion classification by using synthetic speech data. The synthesizer architecture, a variation of Tacotron, generates output frames based on input and is conditioned on an embedding vector containing emotion and speaker information. The vocoder, based on the WaveRNN [148] architecture, generates audio from a Mel-spectrogram [149]. For speech emotion recognition (SER), a CNN-based classifier is used with Mel-frequency cepstral coefficients (MFCCs) as input. The experimental protocol involves training on the Librispeech dataset [78] for TTS and on the Emotional Voices Database (EVD) [150] and Toronto Emotional Speech Set (TESS) [151] for emotional embeddings. The model was evaluated on the Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS) [152] for within-corpus evaluations and on the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [153], Surrey Audio–Visual Expressed Emotion (SAVEE) [154], and Berlin Database of Emotional Speech (EmoDB) [155] for cross-corpus evaluation.

Wang et al. [77] introduced VALL-E (https://www.microsoft.com/en-us/research/project/vall-e-x/, accessed on 14 June 2024), a language model approach for TTS synthesis that utilizes audio codec codes as intermediate representations. VALL-E is trained on a neural codec language model using a large dataset of English speech, enabling personalized speech synthesis with minimal speaker data. The model treats TTS as conditional codec language modeling, leveraging in-context learning capabilities and diverse synthesized outputs in zero-shot scenarios. The training data for VALL-E is sourced from librilight [76]. The speech recognition model is trained on the LibriSpeech dataset and then utilized to transcribe the audio data into phoneme-level alignments. After the hybrid model has been trained, the unlabeled speech data are decoded and transduced to the best phoneme-level alignment paths using a frameshift of 30 ms. The EnCodec model [156] was used to generate the acoustic code matrix for 60,000 h of data, which are used as inputs to the VALL-E neural codec language model for TTS synthesis. A combined dataset consisting of VCTK [112], libriTTS [79], and TTS-Portuguese [110] was used by the authors to determine the baseline for their state-of-the-art zero-shot TTS model called YourTTS [157]. Further, Wang et al. assessed the speaker similarity between the synthesized speech and decompressed enrolled speech using the WavLM-TDNN speaker verification model [158]. They computed the word error rate (WER) in relation to the transcriptions that were originally recorded. The HuBERT-Large [159] model fine-tuned on LibriSpeech was used as the automated speech recognition (ASR) model in the experiment.

Both of the papers reviewed above offer unique approaches to TTS synthesis, each contributing differently to the field with specific methodologies and evaluation strategies. Both the advanced TTS synthesis integrating emotion conditioning and the model using neural codec language modeling for personalized speech generation show demonstrated effectiveness through detailed speaker similarity and naturalness evaluations. However, these papers lack discussions of model interpretability and computational efficiency. Many other works can be found in the literature focusing on various aspects of TTS synthesis, including speaker adaptation, emotion conditioning, and personalized speech generation, contributing to the advancement of TTS technology in different ways [160–163].

3.5. Speech Analysis and Classification

The practice of classifying spoken utterances into predetermined classes or categories is known as speech classification. In order to predict the class or category to which the speech belongs, this is usually accomplished using ML algorithms that examine the acoustic characteristics of the speech signal, such as pitch, intensity, and timing. It is an essential part of numerous speech processing applications, such as language identification, speaker identification, emotion recognition, and voice recognition, and is utilized in different applications such as speech-based customer support systems [164] and automated transcription services [164–167]. There are various methods for classifying speech, including supervised learning, semi-supervised learning, unsupervised learning, and self-supervised learning [168–172]. Speech classification is a challenging problem due to the variability of speech signals, including differences in accent, pronunciation, and speaking style.

For example, Xu [24] focused on speech recognition and pronunciation quality evaluation using DL. The author discussed the process of speech recognition, signal preprocessing, feature extraction, and the development of DL models such as Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs). The study involved 24 college students recording sentences in English for evaluation. The evaluation considered the factors, intonation, speed, rhythm, and tone of voice.

Mu et al. [25] presented a comprehensive Japanese speech evaluation system utilizing a two-layer deep learning model. The system incorporates a Connectionist Temporal Classification (CTC) model for speech-text alignment and segmentation followed by an attention model for word-level speech recognition and evaluation. The study focuses on improving pronunciation accuracy for Japanese language learners, particularly Chinese users, by leveraging a large dataset of learner speech samples. The dataset used in the paper was divided into two sections: first, the announcer's (a native speaker) correct pronunciation along with the pronunciation of an example sentence; second, the pronunciation of words and sentences that users read on a daily basis, with approximately 20,000 Japanese-pronounceable sentences generated daily. Various metrics were used to evaluate model performance, providing insights into the ability to accurately recognize and evaluate pronunciation errors. The model's performance in identifying correct and incorrect phonemes showcased promising results, with low error rates for false recognition and diagnostic recognition errors.

Gong et al. [26] introduced goodness of pronunciation transformer (GoPT), a novel approach to computer-assisted pronunciation training (CAPT) that leverages a transformer self-attention architecture and goodness of pronunciation features. This method enables the simultaneous assessment of multiple aspects of pronunciation quality across the phoneme, word, and utterance levels, including accuracy, fluency, prosody, and stress. The authors employed a public ASR acoustic model trained on the LibriSpeech dataset [78] for their main experiment. Other datasets include TED-LIUM 3 [173], LibriSpeech, and WSJ [174].

The reviewed works focus on speech recognition and pronunciation quality evaluation employing several different approaches and techniques, including the application of DL models for speech recognition and pronunciation quality assessment across the phoneme, word, and utterance levels. These analyses have been applied to various languages, including English, Chinese, and Japanese, and can be used in branding and marketing to evaluate the pronunciation quality of brand names or slogans across different countries and cultures. Many other works can be found in the literature as well, such as [175–177].

Speech Sentiment Analysis

Speech sentiment analysis refers to the process of using techniques to analyze the sentiment or emotional tone expressed in spoken language, such as in conversations, speeches, or recorded audio. This type of sentiment analysis focuses on understanding the speaker's emotions, attitudes, and opinions, and can be used to gain insights into the overall sentiment of the speaker or a group of speakers. Speech sentiment analysis can be challenging due to the complexity and variability of spoken language, including factors such as tone of voice, sarcasm, and cultural nuances. However, recent advancements in artificial intelligence and natural language processing have led to significant improvements in the accuracy and reliability of speech sentiment analysis systems [178,179].

A method for sentiment analysis utilizing characteristics taken from ASR models was presented by Lu et al. [27]. This method uses pretrained ASR characteristics and sentiment decoders to classify sentiment in speech based on two state-of-the-art datasets: the scripted and improvisational interaction IEMOCAP dataset [47] and the natural conversation SWBD-sentiment dataset [105]. The resulting multi-modality model incorporates both audio and ASR transcription as inputs, whereas a single-modality model exclusively employs audio inputs.

Novais et al. [28] discussed a framework for predicting speech emotion using an ensemble of ML methods. The primary focus was on improving social interactions between humans and socially assistive robots, particularly for the elderly, by enhancing robots' ability to recognize and react to human emotions based on speech. Their model introduces an ensemble method that integrates outputs from various pre-existing speech emotion classifiers to predict emotions more accurately. These classifiers analyze audio samples and provide assessments, which are then aggregated to produce a final emotion classification. The methods were validated on RAVDESS [152], SAVEE [154], and TESS [151]. The paper concluded that ensemble methods are a promising avenue for enhancing the performance of emotion recognition systems. The same authors later presented a stack ensemble method that considers both speech and facial expressions to predict emotions [180].

Shon et al. [29] explored two methods for speech sentiment analysis using pretrained language models (LMs). First, a two-step pipeline involves using a pretrained LM as an embedding layer was used for sentiment classification. Second, a semi-supervised end-to-end (E2E) approach incorporating pseudolabels generated by a pretrained LM was used to enhance the sentiment classifier. In the two-step pipeline, the pretrained LM encodes ASR transcripts into a BERT output sequence, which is then classified for sentiment. For the E2E systems, the authors started by creating a baseline system that uses an ASR encoder output as the features for speech sentiment analysis [27]. They suggested a pseudolabel-based semi-supervised training method following this paradigm. The SWBD-Sentiment dataset, which is labeled with three sentiments (positive, neutral, and negative) was used for the tests. In several of their tests, the authors substituted the ASR transcripts from the SWBD-Sentiment and Fisher datasets [71] for the ground truth texts in the two-step process.

The works addressed in this section focus on sentiment analysis in speech using pretrained models. While they share the goal of improving sentiment prediction, they employ different methods and approaches, for instance making use of pretrained ASR characteristics and sentiment decoders (using a two-step pipeline involving encoded ASR transcripts which are subsequently subjected to sentiment classification), employing ensembles of publicly available methods (using pseudolabels produced by a pretrained LM), and more. Many other works can be found in the literature focusing on various aspects of speech sentiment analysis, including emotion recognition, sentiment classification, and social interaction analysis, each contributing to the advancement of speech sentiment analysis technology in different ways [181,182].

3.6. Critical Analysis

This section has presented a review of AI models and techniques which can be used for brand value extraction, text sentiment analysis, semantic analysis, text generation, text-to-speech, speech classification, and speech sentiment analysis. The review covers a wide range of AI models and techniques, including ML, DL, NLP, and speech processing. The applications of these models and techniques has been discussed in the context of various tasks, such as brand value extraction from product descriptions, sentiment analysis of text data, semantic analysis of text, generation of coherent and contextually relevant text, conversion of written text into spoken words, classification of spoken utterances into predetermined classes or categories, and analysis of the sentiment or emotional tone expressed in spoken language. Of course, the applications of these models and techniques are not limited to the tasks discussed in this review, and they can be applied to a wide range of other tasks and domains as well. Further, being such a broad field, there are many other AI models and techniques that can be used for these tasks, and the review presented here is by no means exhaustive. Table 1 provides a summary of the discussed works by category.

4. Datasets

Datasets play a major role in the development and evaluation of AI models and techniques. In this context, high-quality datasets are essential for training, validating, and testing AI models, as they are used to assess the models' performance and generalization capabilities. Table 2 presents a list of datasets used in the discussed works along with a small set identified by the authors, as well as the tasks they were applied to. These datasets cover a wide range of domains and applications, including product descriptions, movie reviews, hotel reviews, newsgroup documents, recognition memory tasks, recall memory tasks, free association norms, concreteness ratings, valence–arousal–dominance ratings, and word memorability. Several of the datasets are publicly available and can be used for developing methods around the aforementioned tasks such as brand value extraction, sentiment analysis, semantic analysis, text generation, text-to-speech, speech classification, and speech sentiment analysis.

5. Conclusions and Future Works

The field of marketing and branding appears to have a future at the convergence of AI and brand name creation and assessment. Artificial intelligence models have proven through the use of natural language processing that they are capable of producing original brand names as well as assessing their possible appeal. However, in certain instances the process may not appear scientifically rigorous. Many examples involve online platforms that generate names for companies, products, and services. These sites likely employ AI models that create brand names by recognizing linguistic patterns and semantic links within existing datasets; of course, the specific algorithms they use are not publicly disclosed. This opens up a field of research that could be explored, as the use of AI in brand name creation and evaluation is currently in its youth.

Therefore, collaboration between AI and human marketers is essential to ensure the creation of compelling and effective brand names. AI models can generate a large volume of potential brand names quickly and efficiently, enabling marketers to explore a wider range of options. This collaboration plots what appears to be a promising future for brand name generation and evaluation. Benefiting from the capabilities of AI in conjunction with human creativity and expertise, marketers can unlock new possibilities for developing impactful and memorable brand names that resonate with consumers.

In this context, AI models can include and analyze consumer sentiment and semantic associations to estimate the potential success of a brand name within a specific market or demographic. However, it is important to note that while at the moment AI can assist in generating and evaluating brand names, the human element remains crucial. One reason for this is that AI models are trained on existing data, and may not always capture the cultural nuances and subtleties that can make a brand name resonate with consumers.

This work tries to provide an overview of AI models and techniques capable of being used for brand value extraction and assessement. From our point of view, and in corroboration with the literature, the applications of these models and techniques have the potential to revolutionize the field of marketing and branding in the future. Techniques include text sentiment analysis, semantic analysis, text generation, text-to-speech, speech classification, and speech sentiment analysis. These techniques are supported in a wide range of AI models, including ML, DL, NLP, and speech processing.

Furthermore, because a major role in the development and evaluation of AI models and techniques is played by data, a section has been dedicated to compiling a list of datasets used in the discussed works along with the tasks they were applied to. These datasets cover a wide range of domains and applications, including product descriptions, movie reviews, hotel reviews, newsgroup documents, recognition memory tasks, recall memory tasks, free association norms, concreteness ratings, valence–arousal–dominance ratings, and word memorability. The list presents a pivotal point for future research, as it provides a starting point for researchers interested in developing methods related to the tasks discussed in this work. It should be noted that the status and availability of these datasets is a key point; while many of them are publicly available and can be used for developing methods for the aforementioned tasks, some are private or no longer available, which may limit the reproducibility of the results obtained in the reviewed works.

Looking ahead, the continued advancement of AI technologies holds great promise for the future of brand name generation and evaluation, particularly in the areas of natural language understanding and sentiment analysis. By leveraging the capabilities of AI in conjunction with human creativity and expertise, marketers can unlock new possibilities for developing impactful and memorable brand names that resonate with consumers.

Future work in the field of AI and branding could explore some of the following directions: (a) the development of AI models that can generate brand names based on specific criteria or requirements, such as target audience or industry, taking into account larger datasets and more complex linguistic and semantic features; (b) the use of AI models for analyzing consumer sentiment and semantic associations to estimate the potential success of a brand name within a specific market or demographic; (c) the integration of AI models with human creativity and expertise; (d) the exploration of cultural nuances and subtleties that can make a brand name resonate with consumers; (e) the development of additional AI models that can evaluate the memorability and brandability of brand names; and (f) the creation of dedicated datasets that can be used to train and evaluate AI models for brand name generation and evaluation. This last point is particularly relevant, as from the first instant the dataset should take into account the specific requirements and constraints of the branding industry, such as the need for unique, memorable, and impactful brand names, the importance of cultural and linguistic considerations, the age groups and demographics of the target audience, and the industry or market segment in which the brand operates. Additionally, datasets should be large, diverse, and unbiased, well annotated, and labeled in order to ensure the quality and reliability of AI models trained on them. Finally, whenever possible datasets should be publicly available and well documented to ensure reproducibility and transparency in the development and evaluation of the resulting AI models.

Author Contributions: Conceptualization, P.J.S.C. and J.M.F.R.; methodology, P.J.S.C. and J.M.F.R.; software, M.L.; validation, M.L., P.J.S.C. and J.M.F.R.; formal analysis, M.L., P.J.S.C. and J.M.F.R.; investigation, M.L., P.J.S.C. and J.M.F.R.; resources, P.J.S.C. and J.M.F.R.; data curation, M.L.; writing—original draft preparation, M.L., P.J.S.C. and J.M.F.R.; writing—review and editing, M.L., P.J.S.C. and J.M.F.R.; visualization, M.L., P.J.S.C. and J.M.F.R.; supervision, P.J.S.C.; project administration, P.J.S.C.; funding acquisition, P.J.S.C. and J.M.F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020, accessed on 23 June 2024) and ref. UIDP/04516/2020 (https://doi.org/10.54499/UIDP/04516/2020, accessed on 23 June 2024) with the financial support of FCT.IP.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Arora, S.; Kalro, A.D.; Sharma, D. A comprehensive framework of brand name classification. *J. Brand Manag.* **2015**, *22*, 79–116. [CrossRef]
2. Moro Visconti, R. Domain Name Valuation: Internet Traffic Monetization and IT Portfolio Bundling. 2017. Available online: https://ssrn.com/abstract=3028534 (accessed on 14 June 2024).
3. Eskiev, M. Naming as one of the most important elements of brand management. *SHS Web Conf.* **2021**, *128*, 01028.

4. Sabeh, K.; Kacimi, M.; Gamper, J. OpenBrand: Open Brand Value Extraction from Product Descriptions. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5), Dublin, Ireland, 26, May 2022*; Malmasi, S., Rokhlenko, O., Ueffing, N., Guy, I., Agichtein, E., Kallumadi, S., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 161–170. [CrossRef]
5. Sabeh, K.; Kacimi, M.; Gamper, J. GAVI: A Category-Aware Generative Approach for Brand Value Identification. In Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023), Virtual, 16–17 December 2023; pp. 110–119.
6. Nurmambetov, D.; Dauylov, S.; Bogdanchikov, A. Kazakh Names Generator Using Deep Learning. *Her. Kazakh-Br. Tech. Univ.* **2021**, *17*, 171–177.
7. Muhammad, P.F.; Kusumaningrum, R.; Wibowo, A. Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. *Procedia Comput. Sci.* **2021**, *179*, 728–735. [CrossRef]
8. Jiang, H.; Hu, C.; Jiang, F. Text Sentiment Analysis of Movie Reviews Based on Word2Vec-LSTM. In Proceedings of the 14th International Conference on Advanced Computational Intelligence (ICACI), Wuhan, China, 22 July 2022; pp. 129–134.
9. Liu, Y. Transgender Community Sentiment Analysis from Social Media Data: A Natural Language Processing Approach. *Gen. Surgery Clin. Med.* **2023**, *1*, 127–131.
10. Hassan, J.; Shoaib, U. Multi-class review rating classification using deep recurrent neural network. *Neural Process. Lett.* **2020**, *51*, 1031–1048. [CrossRef]
11. Mendes, G.A.; Martins, B. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Proceedings of the European Conference on Information Retrieval, Dublin, Ireland, 2–6 April 2023*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 84–100.
12. Wang, X.; Dong, X.; Chen, S. Text duplicated-checking algorithm implementation based on natural language semantic analysis. In Proceedings of the IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 5 June 2020; pp. 732–735.
13. Maksutov, A.A.; Zamyatovskiy, V.I.; Vyunnikov, V.N.; Kutuzov, A.V. Knowledge base collecting using natural language processing algorithms. In Proceedings of the IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg/Moscow, Russia, 27–30 January 2020; pp. 405–407.
14. Li, W. Analysis of semantic comprehension algorithms of natural language based on robot's questions and answers. In Proceedings of the IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 25 August 2020; pp. 1021–1024.
15. Aka, A.; Bhatia, S.; McCoy, J. Semantic determinants of memorability. *Cognition* **2023**, *239*, 105497. [CrossRef] [PubMed]
16. Tuckute, G.; Mahowald, K.; Isola, P.; Oliva, A.; Gibson, E.; Fedorenko, E. Intrinsically memorable words have unique associations with their meanings. *PsyArXiv* **2022**. [CrossRef]
17. Sadhuram, M.V.; Soni, A. Natural language processing based new approach to design factoid question answering system. In Proceedings of the 2nd International Conference on Inventive Research in Computing Applications (ICIRCA), Virtual, 15–17 July 2020; pp. 276–281.
18. Song, Z.; Zheng, X.; Liu, L.; Xu, M.; Huang, X.J. Generating responses with a specific emotion in dialog. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3685–3695.
19. Liu, Y.; Lin, Z.; Liu, F.; Dai, Q.; Wang, W. Generating paraphrase with topic as prior knowledge. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 2381–2384.
20. He, X. Parallel Refinements for Lexically Constrained Text Generation with BART. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 8653–8666. [CrossRef]
21. Wang, Y.; Wood, I.; Wan, S.; Dras, M.; Johnson, M. Mention Flags (MF): Constraining Transformer-based Text Generators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 2–4 August 2021*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistic: Stroudsburg, PA, USA, 2021; pp. 103–113. [CrossRef]
22. Latif, S.; Shahid, A.; Qadir, J. Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation. *Appl. Acoust.* **2023**, *210*, 109425. [CrossRef]
23. Rashid, M.; Priya; Singh, H. Text to speech conversion in Punjabi language using nourish forwarding algorithm. *Int. J. Inf. Tecnol.* **2022**, *14*, 559–568. [CrossRef]
24. Xu, Y. English speech recognition and evaluation of pronunciation quality using deep learning. *Mob. Inf. Syst.* **2022**, *2022*, 1–12. [CrossRef]
25. Mu, D.; Sun, W.; Xu, G.; Li, W. Japanese Pronunciation Evaluation Based on DDNN. *IEEE Access* **2020**, *8*, 218644–218657. [CrossRef]
26. Gong, Y.; Chen, Z.; Chu, I.H.; Chang, P.; Glass, J. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 7262–7266.

27. Lu, Z.; Cao, L.; Zhang, Y.; Chiu, C.C.; Fan, J. Speech sentiment analysis via pre-trained features from end-to-end asr models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), virtual-Barcelona, Spain, 4–8 May 2020; pp. 7149–7153.
28. Novais, R.; Cardoso, P.J.S.; Rodrigues, J.M.F. Emotion classification from speech by an ensemble strategy. In Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, Lisboa, Portugal, 31 August–2 September, 2022; pp. 85–90.
29. Shon, S.; Brusco, P.; Pan, J.; Han, K.J.; Watanabe, S. Leveraging Pre-trained Language Model for Speech Sentiment Analysis. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, Brno, Czechia, 30 August–3 September 2021*; International Speech Communication Association: Brno, Czech Republic, 2021; pp. 566–570.
30. Sabeh, K. Open-Brand: The Dataset Contains over 250 K Product Brand-Value Annotations with More Than 50 k Unique Values across Eight Main Categories of Amazon Product Profiles. 2022. Available online: https://github.com/kassemsabeh/open-brand (accessed on 13 March 2024).
31. Ni, J.; Li, J.; McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 188–197.
32. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
33. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40. [CrossRef]
34. Caschera, M.C.; Grifoni, P.; Ferri, F. Emotion Classification from Speech and Text in Videos Using a Multimodal Approach. *Multimodal Technol. Interact.* **2022**, *6*, 28. [CrossRef]
35. Bogdanchikov, A.; Ayazbayev, D.; Varlamis, I. Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. *Big Data Cogn. Comput.* **2022**, *6*, 123. [CrossRef]
36. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
37. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.J.; McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60.
38. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space (2013). *arXiv* **2023**, arXiv:1301.3781.
39. Khyani, D.; Siddhartha, B.; Niveditha, N.; Divya, B. An interpretation of lemmatization and stemming in natural language processing. *J. Univ. Shanghai Sci. Technol.* **2021**, *22*, 350–357.
40. Pramana, R.; Subroto, J.J.; Gunawan, A.A.S. Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity. In Proceedings of the IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), Yogyakarta, Indonesia, 4–5 November 2022; pp. 1–6.
41. Maas, A.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 19–24 June 2011; pp. 142–150.
42. Datafiniti. Hotel Reviews. 2019. Available online: https://www.kaggle.com/datasets/datafiniti/hotel-reviews (accessed on 13 March 2024).
43. Lakshmipathi, N. IMDB Dataset of 50 K Movie Reviews. 2019. Available online: https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (accessed on 13 March 2024).
44. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
45. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8440–8451. [CrossRef]
46. Citron, F.M.; Lee, M.; Michaelis, N. Affective and psycholinguistic norms for German conceptual metaphors (COMETA). *Behav. Res. Methods* **2020**, *52*, 1056–1072. [CrossRef] [PubMed]
47. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
48. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. In *Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Glasgow, UK, 2015; Volume 28.
49. Datafiniti. Consumer Reviews of Amazon Products. 2019. Available online: https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products (accessed on 13 March 2024).
50. Fairfield, B.; Ambrosini, E.; Mammarella, N.; Montefinese, M. Affective norms for Italian words in older adults: Age differences in ratings of valence, arousal and dominance. *PLoS ONE* **2017**, *12*, e0169472. [CrossRef]

51. Soares, A.P.; Comesaña, M.; Pinheiro, A.P.; Simões, A.; Frade, C.S. The adaptation of the Affective Norms for English words (ANEW) for European Portuguese. *Behav. Res. Methods* **2012**, *44*, 256–269. [CrossRef] [PubMed]
52. Schmidtke, D.S.; Schröder, T.; Jacobs, A.M.; Conrad, M. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behav. Res. Methods* **2014**, *46*, 1108–1118. [CrossRef]
53. Imbir, K.K. The Affective Norms for Polish Short Texts (ANPST) database properties and impact of participants' population and sex on affective ratings. *Front. Psychol.* **2017**, *8*, 251141. [CrossRef] [PubMed]
54. Imbir, K.K. Affective norms for 4900 Polish words reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Front. Psychol.* **2016**, *7*, 174568. [CrossRef]
55. Võ, M.L.; Conrad, M.; Kuchinke, L.; Urton, K.; Hofmann, M.J.; Jacobs, A.M. The Berlin affective word list reloaded (BAWL-R). *Behav. Res. Methods* **2009**, *41*, 534–538. [CrossRef]
56. Aka, A.; Phan, T.D.; Kahana, M.J. Predicting recall of words and lists. *J. Exp. Psychol. Learn. Mem. Cogn.* **2021**, *47*, 765. [CrossRef]
57. Brysbaert, M.; Warriner, A.B.; Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **2014**, *46*, 904–911. [CrossRef]
58. Ćoso, B.; Guasch, M.; Ferré, P.; Hinojosa, J.A. Affective and concreteness norms for 3022 Croatian words. *Q. J. Exp. Psychol.* **2019**, *72*, 2302–2312. [CrossRef] [PubMed]
59. Xie, H.; Lin, W.; Lin, S.; Wang, J.; Yu, L.C. A multi-dimensional relation model for dimensional sentiment analysis. *Inf. Sci.* **2021**, *579*, 832–844. [CrossRef]
60. Yu, L.C.; Lee, L.H.; Hao, S.; Wang, J.; He, Y.; Hu, J.; Lai, K.R.; Zhang, X. Building Chinese affective resources in valence-arousal dimensions. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 540–545.
61. Lee, L.H.; Li, J.H.; Yu, L.C. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–18. [CrossRef]
62. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. Dbpedia—A large-scale, multilingual knowledge base extracted from wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
63. Eilola, T.M.; Havelka, J. Affective norms for 210 British English and Finnish nouns. *Behav. Res. Methods* **2010**, *42*, 134–140. [CrossRef] [PubMed]
64. Buechel, S.; Hahn, U. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv* **2022**, arXiv:2205.01996.
65. Buechel, S.; Hahn, U. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In Proceedings of the 11th Linguistic Annotation Workshop, Valencia, Spain, 3 April 2017; pp. 1–12.
66. Francisco, V.; Hervás, R.; Peinado, F.; Gervás, P. EmoTales: Creating a corpus of folk tales with emotional annotations. *Lang. Resour. Eval.* **2012**, *46*, 341–381. [CrossRef]
67. Loza Mencía, E.; Fürnkranz, J. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Antwerp, Belgium, 15–19 September 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 50–65.
68. Preoţiuc-Pietro, D.; Schwartz, H.A.; Park, G.; Eichstaedt, J.; Kern, M.; Ungar, L.; Shulman, E. Modelling valence and arousal in facebook posts. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, San Diego, CA, USA, 16 June 2016; pp. 9–15.
69. Monnier, C.; Syssau, A. Affective norms for French words (FAN). *Behav. Res. Methods* **2014**, *46*, 1128–1137. [CrossRef] [PubMed]
70. Gilet, A.L.; Grühn, D.; Studer, J.; Labouvie-Vief, G. Valence, arousal, and imagery ratings for 835 French attributes by young, middle-aged, and older adults: The French Emotional Evaluation List (FEEL). *Eur. Rev. Appl. Psychol.* **2012**, *62*, 173–181. [CrossRef]
71. Cieri, C.; Miller, D.; Walker, K. The Fisher corpus: A resource for the next generations of speech-to-text. In Proceedings of the LREC, Lisbon, Portugal, 26–28 May 2004; Volume 4, pp. 69–71.
72. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
73. Wallace, B.C.; Kertz, L.; Charniak, E. Humans require context to infer ironic intent (so computers probably do, too). In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 23–25 June 2014; pp. 512–516.
74. Kapucu, A.; Kılıç, A.; Özkılıç, Y.; Sarıbaz, B. Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychol. Rep.* **2021**, *124*, 188–209. [CrossRef] [PubMed]
75. Kanske, P.; Kotz, S.A. Leipzig affective norms for German: A reliability study. *Behav. Res. Methods* **2010**, *42*, 987–991. [CrossRef]
76. Kahn, J.; Riviere, M.; Zheng, W.; Kharitonov, E.; Xu, Q.; Mazaré, P.E.; Karadayi, J.; Liptchinsky, V.; Collobert, R.; Fuegen, C. Libri-light: A benchmark for asr with limited or no supervision. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7669–7673.
77. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv* **2023**, arXiv:2301.02111.

78. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. LibriSpeech: An ARS corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
79. Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R.J.; Jia, Y.; Chen, Z.; Wu, Y. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1526–1530. [CrossRef]
80. Pinheiro, A.P.; Dias, M.; Pedrosa, J.; Soares, A.P. Minho Affective Sentences (MAS): Probing the roles of sex, mood, and empathy in affective ratings of verbal stimuli. *Behav. Res. Methods* **2017**, *49*, 698–716. [CrossRef]
81. Moors, A.; De Houwer, J.; Hermans, D.; Wanmaker, S.; Van Schie, K.; Van Harmelen, A.L.; De Schryver, M.; De Winne, J.; Brysbaert, M. Norms of valence, arousal, dominance, and age of acquisition for 4300 Dutch words. *Behav. Res. Methods* **2013**, *45*, 169–177. [CrossRef] [PubMed]
82. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.; Potts, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
83. Deng, L.; Wiebe, J. MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015*; Mihalcea, R., Chai, J., Sarkar, A., Eds.; Curran Associates: New York, NY, USA, 2015; pp. 1323–1328. [CrossRef]
84. Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L. Ms marco: A human-generated machine reading comprehension dataset. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
85. Dolan, W.; Quirk, C.; Brockett, C.; Dolan, B. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, 23 - 27 August 2004; pp. 350–356.
86. Williams, A.; Nangia, N.; Bowman, S.R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv* **2017**, arXiv:1704.05426.
87. Riegel, M.; Wierzba, M.; Wypych, M.; Żurawski, Ł.; Jednoróg, K.; Grabowska, A.; Marchewka, A. Nencki affective word list (NAWL): The cultural adaptation of the Berlin affective word list–reloaded (BAWL-R) for Polish. *Behav. Res. Methods* **2015**, *47*, 1222–1236. [CrossRef] [PubMed]
88. Mohammad, S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 174–184.
89. Ohsumed. XmdvTool Home Page: Downloads. 2005. Available online: http://davis.wpi.edu/xmdv/datasets/ohsumed.html (accessed on 13 March 2024).
90. Chelba, C.; Mikolov, T.; Schuster, M.; Ge, Q.; Brants, T.; Koehn, P.; Robinson, T. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. 2014. Available online: https://arxiv.org/abs/1312.3005 (accessed on 13 March 2024).
91. Citron, F.M.; Cacciari, C.; Kucharski, M.; Beck, L.; Conrad, M.; Jacobs, A.M. When emotions are expressed figuratively: Psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). *Behav. Res. Methods* **2016**, *48*, 91–111. [CrossRef] [PubMed]
92. Lu, Z. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. 2011. Available online: https://pubmed.ncbi.nlm.nih.gov/21245076/ (accessed on 13 March 2024).
93. Iyer, S.; Dandekar, N.; Csernai, K. First Quora Dataset Release: Question Pairs—Data @ Quora—Quora. 2012. Available online: https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs (accessed on 13 March 2024).
94. Kahana, M.J.; Aggarwal, E.V.; Phan, T.D. The variability puzzle in human memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **2018**, *44*, 1857. [CrossRef] [PubMed]
95. Healey, M.K.; Crutchley, P.; Kahana, M.J. Individual differences in memory search and their relation to intelligence. *J. Exp. Psychol. Gen.* **2014**, *143*, 1553. [CrossRef] [PubMed]
96. Lohnas, L.J.; Kahana, M.J. Parametric effects of word frequency in memory for mixed frequency lists. *J. Exp. Psychol. Learn. Mem. Cogn.* **2013**, *39*, 1943. [CrossRef] [PubMed]
97. Thoma, M. The Reuters Dataset. 2017. Available online: https://martin-thoma.com/nlp-reuters (accessed on 13 March 2024).
98. Zhang, J.; Zhang, Z.; Wang, Y.; Yan, Z.; Song, Q.; Huang, Y.; Li, K.; Povey, D.; Wang, Y. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. *arXiv* **2021**, arXiv:2104.01378.
99. Stadthagen-Gonzalez, H.; Imbault, C.; Pérez Sánchez, M.A.; Brysbaert, M. Norms of valence and arousal for 14,031 Spanish words. *Behav. Res. Methods* **2017**, *49*, 111–123. [CrossRef]
100. Marelli, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; Zamparelli, R. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 1–8.
101. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. *arXiv* **2015**, arXiv:1508.05326.

102. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune bert for text classification? In Proceedings of the Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, 18–20 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 194–206.
103. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphia, PA, USA, 6 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 79–86. [CrossRef]
104. Brysbaert, M.; New, B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* **2009**, *41*, 977–990. [CrossRef] [PubMed]
105. Chen, E.; Lu, Z.; Xu, H.; Cao, L.; Zhang, Y.; Fan, J. A large scale speech sentiment corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6549–6555.
106. Scott, G.G.; Keitel, A.; Becirspahic, M.; Yao, B.; Sereno, S.C. The Glasgow Norms: Ratings of 5500 words on nine scales. *Behav. Res. Methods* **2019**, *51*, 1258–1270. [CrossRef] [PubMed]
107. Report, N. The 500 Largest Firms in the World Rated by Net Effect. 2021. Available online: https://netimpactreport.com/datasets/largest-500 (accessed on 13 March 2024).
108. GauravArora1091. Top 100 Global Brands by Brandirectory-2022. 2022. Available online: https://www.kaggle.com/datasets/gauravarora1091/top-100-global-brands-by-brandirectory2022 (accessed on 13 March 2024).
109. Li, X.; Roth, D. Learning question classifiers. In Proceedings of the COLING: The 19th International Conference on Computational Linguistics, Taipei, Taiwan, 24 August–1 September 2002.
110. Casanova, E.; Junior, A.C.; Shulby, C.; Oliveira, F.S.d.; Teixeira, J.P.; Ponti, M.A.; Aluísio, S. TTS-Portuguese Corpus: A corpus for speech synthesis in Brazilian Portuguese. *Lang. Resour. Eval.* **2022**, *56*, 1043–1055. [CrossRef]
111. Nelson, D.L.; McEvoy, C.L.; Schreiber, T.A. The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods Instruments Comput.* **2004**, *36*, 402–407. [CrossRef] [PubMed]
112. Veaux, C.; Yamagishi, J.; MacDonald, K. Superseded-Cstr Vctk Corpus: English Multi-Speaker Corpus for Cstr Voice Cloning Toolkit The Centre for Speech Technology Research (CSTR), University of Edinburgh. 2016. Available online: https://datashare.ed.ac.uk/handle/10283/3443 (accessed on 14 June 2024).
113. Verheyen, S.; De Deyne, S.; Linsen, S.; Storms, G. Lexicosemantic, affective, and distributional norms for 1000 Dutch adjectives. *Behav. Res. Methods* **2020**, *52*, 1108–1121. [CrossRef]
114. Warriner, A.B.; Kuperman, V.; Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **2013**, *45*, 1191–1207. [CrossRef]
115. Kowsari, K.; Brown, D.E.; Heidarysafa, M.; Meimandi, K.J.; Gerber, M.S.; Barnes, L.E. Hdltex: Hierarchical deep learning for text classification. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 364–371.
116. Xu, X.; Li, J.; Chen, H. Valence and arousal ratings for 11,310 simplified Chinese words. *Behav. Res. Methods* **2022**, *54*, 26–41. [CrossRef] [PubMed]
117. Yee, L.T. Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in Cantonese speakers in Hong Kong. *PLoS ONE* **2017**, *12*, e0174569. [CrossRef] [PubMed]
118. Yelp, I. Yelp Dataset. 2022. Available online: https://www.kaggle.com/yelp-dataset/yelp-dataset (accessed on 13 March 2024).
119. Labs, P.D. 7+ Million Company Dataset. 2019. Available online: https://www.kaggle.com/datasets/peopledatalabssf/free-7-million-company-dataset (accessed on 13 March 2024).
120. Lang, K. 20 Newsgroups. 2008. Available online: http://qwone.com/~jason/20Newsgroups/ (accessed on 13 March 2024).
121. Lilleberg, J.; Zhu, Y.; Zhang, Y. Support vector machines and word2vec for text classification with semantic features. In Proceedings of the IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), Beijing, China, 6–8 July 2015; pp. 136–140.
122. Labs, P.D. Company Data to Get Intelligence on 22.9+ Million Companies. 2024. Available online: https://www.peopledatalabs.com/company-dataset (accessed on 13 March 2024).
123. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
124. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 517,
125. Cox, G.E.; Hemmer, P.; Aue, W.R.; Criss, A.H. Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *J. Exp. Psychol. Gen.* **2018**, *147*, 545. [CrossRef] [PubMed]
126. Mahowald, K.; Isola, P.; Fedorenko, E.; Gibson, E.; Oliva, A. Memorable Words Are Monogamous: The Role of Synonymy and Homonymy in Word Recognition Memory. Preprint at PsyArxiv. 2018. Available online: https://psyarxiv.com/p6kv9 (accessed on 14 June 2024).
127. Maulud, D.H.; Zeebaree, S.R.; Jacksi, K.; Sadeeq, M.A.M.; Sharif, K.H. State of art for semantic analysis of natural language processing. *Qubahan Acad. J.* **2021**, *1*, 21–28. [CrossRef]
128. Doyal, A.S.; Sender, D.; Nanda, M.; Serrano, R.A. ChatGPT and artificial intelligence in medical writing: Concerns and ethical considerations. *Cureus* **2023**, *15*, e43292.

129. Iqbal, T.; Qureshi, S. The survey: Text generation models in deep learning. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 2515–2528. [CrossRef]
130. Yu, W.; Zhu, C.; Li, Z.; Hu, Z.; Wang, Q.; Ji, H.; Jiang, M. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.* **2022**, *54*, 1–38. [CrossRef]
131. Rajpurkar, P.; Jia, R.; Liang, P. Know what you don't know: Unanswerable questions for SQuAD. *arXiv* **2018**, arXiv:1806.03822.
132. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems, 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Glasgow, UK, 2014; Volume 27.
133. Shen, X.; Su, H.; Li, W.; Klakow, D. NEXUS Network: Connecting the Preceding and the Following in Dialogue Generation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; Association for Computational Linguistics (ACL): New York, NY, USA, 2018; pp. 4316–4327. [CrossRef]
134. Wang, Y.; Zhao, X.; Zhao, D. Overview of the NLPCC 2022 shared task: Multi-modal dialogue understanding and generation. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Beijing, China, 22–23 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 328–335.
135. Lan, W.; Qiu, S.; He, H.; Xu, W. A continuously growing dataset of sentential paraphrases. *arXiv* **2017**, arXiv:1708.00391.
136. Mou, L.; Yan, R.; Li, G.; Zhang, L.; Jin, Z. Backward and forward language modeling for constrained sentence generation. *arXiv* **2015**, arXiv:1512.06612.
137. Hokamp, C.; Liu, Q. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017.
138. Miao, N.; Zhou, H.; Mou, L.; Yan, R.; Li, L. CGMH: Constrained sentence generation by metropolis-hastings sampling. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6834–6842.
139. Zhang, Y.; Wang, G.; Li, C.; Gan, Z.; Brockett, C.; Dolan, W. POINTER: Constrained Progressive Text Generation via Insertion-based Generative Pre-training. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 8649–8670.
140. He, X.; Li, V. Show me how to revise: Improving lexically constrained sentence generation with XLNet. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2–9 February 2021; Volume 35, pp. 12989–12997.
141. Lin, B.Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; Ren, X. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2020; pp. 1823–1840. [CrossRef]
142. Dušek, O.; Novikova, J.; Rieser, V. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Comput. Speech Lang.* **2020**, *59*, 123–156. [CrossRef]
143. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8948–8957.
144. Lorusso, M.L.; Borasio, F.; Panetto, P.; Curioni, M.; Brotto, G.; Pons, G.; Carsetti, A.; Molteni, M. Validation of a Web App Enabling Children with Dyslexia to Identify Personalized Visual and Auditory Parameters Facilitating Online Text Reading. *Multimodal Technol. Interact.* **2024**, *8*, 5. [CrossRef]
145. Abdulrahman, A.; Richards, D. Is Natural Necessary? Human Voice versus Synthetic Voice for Intelligent Virtual Agents. *Multimodal Technol. Interact.* **2022**, *6*, 51. [CrossRef]
146. Pathak, A.; Velasco, C.; Spence, C. The sound of branding: An analysis of the initial phonemes of popular brand names. *J. Brand Manag.* **2020**, *27*, 339–354. [CrossRef]
147. Vidal-Mestre, M.; Freire-Sánchez, A.; Calderón-Garrido, D.; Faure-Carvallo, A.; Gustems-Carnicer, J. Audio identity in branding and brand communication strategy: A systematic review of the literature on audio branding. *Prof. Inf./Inf. Prof.* **2022**, *31*. [CrossRef]
148. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; van den Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient Neural Audio Synthesis. In Proceedings of the International Conference on Machine Learning 2018, Stockholm, Sweden, 10–15 July 2018.
149. Roberts, L. Understanding the mel spectrogram. *Medium* **2024**. Available online: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53 (accessed on 13 June 2024).
150. Adigwe, A.; Tits, N.; Haddad, K.E.; Ostadabbas, S.; Dutoit, T. The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems. *arXiv* **2018**, arXiv:1806.09514.
151. Dupuis, K.; Pichora-Fuller, M.K. *Toronto Emotional Speech Set (TESS)*; University of Toronto, Psychology Department: Toronto, ON, Canada, 2010.
152. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]

153. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef] [PubMed]
154. Jackson, P.; Haq, S. *Surrey Audio-Visual Expressed Emotion (Savee) Database*; University of Surrey: Guildford, UK, 2014.
155. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; Volume 5, pp. 1517–1520.
156. Défossez, A.; Copet, J.; Synnaeve, G.; Adi, Y. High Fidelity Neural Audio Compression. *Trans. Mach. Learn. Res.* **2023**, *36*.
157. Casanova, E.; Weber, J.; Shulby, C.D.; Junior, A.C.; Gölge, E.; Ponti, M.A. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 2709–2720.
158. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
159. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
160. Bang, C.W.; Chun, C. Effective Zero-Shot Multi-Speaker Text-to-Speech Technique Using Information Perturbation and a Speaker Encoder. *Sensors* **2023**, *23*, 9591. [CrossRef]
161. Alonso Martin, F.; Malfaz, M.; Castro-Gonzalez, A.; Castillo, J.C.; Salichs, M.A. Four-Features Evaluation of Text to Speech Systems for Three Social Robots. *Electronics* **2020**, *9*, 267. [CrossRef]
162. Ning, Y.; He, S.; Wu, Z.; Xing, C.; Zhang, L.J. A Review of Deep Learning Based Speech Synthesis. *Appl. Sci.* **2019**, *9*, 4050. [CrossRef]
163. Nazir, O.; Malik, A. Deep Learning End to End Speech Synthesis: A Review. In Proceedings of the 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), Delhi, India, 21–23 May 2021; pp. 66–71. [CrossRef]
164. Han, W.; Jiang, T.; Li, Y.; Schuller, B.; Ruan, H. Ordinal learning for emotion recognition in customer service calls. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, Barcelona, 4–8 May 2020; pp. 6494–6498.
165. Fox, C.B.; Israelsen-Augenstein, M.; Jones, S.; Gillam, S.L. An evaluation of expedited transcription methods for school-age children's narrative language: Automatic speech recognition and real-time transcription. *J. Speech Lang. Hear. Res.* **2021**, *64*, 3533–3548. [CrossRef] [PubMed]
166. Ling, S.; Liu, Y.; Salazar, J.; Kirchhoff, K. Deep contextualized acoustic representations for semi-supervised speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), virtual, 4–8 May 2020; pp. 6429–6433.
167. Kim, Y.; Levy, J.; Liu, Y. Speech sentiment and customer satisfaction estimation in socialbot conversations. *arXiv* **2020**, arXiv:2008.12376.
168. Singh, A.; Anand, R. Speech Recognition Using Supervised and Unsupervised Learning Techniques. In Proceedings of the International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, MP, India, 12–14 December 2015; pp. 691–696. [CrossRef]
169. Khonglah, B.; Madikeri, S.; Dey, S.; Bourlard, H.; Motlicek, P.; Billa, J. Incremental semi-supervised learning for multi-genre speech recognition. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, Barcelona, 4–8 May 2020; pp. 7419–7423.
170. Baevski, A.; Hsu, W.N.; Conneau, A.; Auli, M. Unsupervised speech recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 27826–27839.
171. Lin, G.T.; Hsu, C.J.; Liu, D.R.; Lee, H.Y.; Tsao, Y. Analyzing the robustness of unsupervised speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 8202–8206.
172. Yue, X.; Li, H. Phonetically Motivated Self-Supervised Speech Representation Learning. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 746–750.
173. Hernandez, F.; Nguyen, V.; Ghannay, S.; Tomashenko, N.; Esteve, Y. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Proceedings of the Speech and Computer: 20th International Conference, SPECOM 2018, Proceedings 20, Leipzig, Germany, 18–22 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 198–208.
174. Marcus, M.; Santorini, B.; Marcinkiewicz, M.A. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
175. Tivarekar, R.P.; Khadye, R.M.; Chavande, S.R.; Talkatkar, P.S. Review of Deep Speech Recognizer using Transcriber. In Proceedings of the 6th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 8–9 December 2023; pp. 460–463.
176. Alharbi, S.; Alrazgan, M.; Alrashed, A.; Alnomasi, T.; Almojel, R.; Alharbi, R.; Alharbi, S.; Alturki, S.; Alshehri, F.; Almojil, M. Automatic speech recognition: Systematic literature review. *IEEE Access* **2021**, *9*, 131858–131876. [CrossRef]
177. Mohamed, A.; Lee, H.y.; Borgholt, L.; Havtorn, J.D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.W.; Livescu, K.; Maaløe, L.; et al. Self-supervised speech representation learning: A review. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1179–1210. [CrossRef]
178. Kumar, T.; Mahrishi, M.; Nawaz, S. A review of speech sentiment analysis using machine learning. In Proceedings of the Trends in Electronics and Health Informatics: TEHI 2021, Kanpur, India, 16–17 December 2021; pp. 21–28.

179. Maghilnan, S.; Kumar, M.R. Sentiment analysis on speaker specific speech data. In Proceedings of the International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 23–24 June 2017; pp. 1–5.
180. Cardoso, P.J.S.; Rodrigues, J.M.F.; Novais, R. Multimodal Emotion Classification Supported in the Aggregation of Pre-trained Classification Models. In Proceedings of the International Conference on Computational Science, Prague, Czechia, 3–5 July 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 433–447.
181. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
182. Das, R.; Singh, T.D. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.