Tackling Sexist Hate Speech: Cross-Lingual Detection and Multilingual Insights from Social Media

---

By
Aiqi Jiang

A Dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

School of Electronic Engineering and Computer Science
of the
Queen Mary University of London

Supervisor: Dr. Arkaitz Zubiaga

London, UK

June 2024

Abstract

With the widespread use of social media, the proliferation of online communication presents both opportunities and challenges for fostering a respectful and inclusive digital environment. Due to the anonymity and weak regulations of social media platforms, the rise of hate speech has become a significant concern, particularly against specific individuals or groups based on race, religion, ethnicity, or gender, posing a severe threat to human rights. Sexist hate speech is a prevalent form of online hate that often manifests itself through gender-based violence and discrimination, challenging societal norms and legal systems. Despite the advances in natural language processing techniques for detecting offensive and sexist content, most research still focuses on monolingual (primarily English) contexts, neglecting the multilingual nature of online platforms. This gap highlights the need for effective and scalable strategies to address the linguistic diversity and cultural variations in hate speech. Cross-language transfer learning and state-of-the-art multilingual pre-trained language models provide potential solutions to improve the detection efficiency of low-resource languages by leveraging data from high-resource languages. Additional knowledge is crucial to facilitate the models' performance in detecting culturally varying expressions of sexist hate speech in different languages.

In this thesis, we delve into the complex area of identifying sexist hate speech in social media across diverse languages pertaining to different language families, with a focus on sexism and a broad exploration of datasets, methodologies, and barriers inherent in mitigating online hate speech in cross-lingual and multilingual scenarios. We primarily apply cross-lingual transfer learning techniques to detect sexist hate

speech, aiming to leverage knowledge acquired from related linguistic data in order to improve performance in a target language. We also investigate the integration of external knowledge to deepen the understanding of sexism in multilingual social media contexts, addressing both the challenges of linguistic diversity and the need for comprehensive, culturally sensitive hate speech detection models.

Specifically, it embarks on a comprehensive survey of tackling cross-lingual hate speech online, summarising existing datasets and cross-lingual approaches, as well as highlighting challenges and frontiers in this field. It then presents a first contribution to the field, the creation of the Sina Weibo Sexism Review (SWSR) dataset in Chinese —a pioneering resource that not only fills a crucial gap in limited resources but also lays the foundation for relevant cross-lingual investigations. Additionally, it examines how cross-lingual techniques can be utilised to generate domain-aware word embeddings, and explores the application of these embeddings in a cross-lingual hate speech framework, thereby enhancing the capacity to capture the subtleties of sexist hate speech across diverse languages. Recognising the significance of linguistic nuances in multilingual and cross-lingual settings, another innovation consists in proposing and evaluating a series of multilingual and cross-lingual models tailored for detecting sexist hate speech. By leveraging the capacity of shared knowledge and features across languages, these models significantly advance the state-of-the-art in identifying online sexist hate speech. As societies continue to deal with the complexities of social media, the findings and methodologies presented in this thesis could effectively help foster more inclusive and respectful online content across languages.

Publications

Research Papers from this Thesis

1. **Aiqi Jiang**, Arkaitz Zubiaga. Cross-lingual Offensive Language Detection: A Systematic Review of Datasets, Transfer Approaches and Challenges. (Submitted to the journal ACM Computing Surveys and under review)

2. **Aiqi Jiang**, Arkaitz Zubiaga. SexWEs: Domain-Aware Word Embeddings via Cross-lingual Semantic Specialisation for Chinese Sexism Detection in Social Media. In Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM 2023).

3. **Aiqi Jiang**, Xiaohan Yang, Yang Liu, Arkaitz Zubiaga. SWSR: A Chinese dataset and lexicon for online sexism detection, Online Social Networks and Media, Volume 27, 2022, 100182, ISSN 2468-6964.

4. **Aiqi Jiang**, Arkaitz Zubiaga. Cross-lingual Capsule Network for Hate Speech Detection in Social Media. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21). Association for Computing Machinery, New York, NY, USA, 217–223, 2021.

5. **Aiqi Jiang**. QMUL-NLP at HASOC 2019: Offensive Content Detection and Classification in Social Media. FIRE (Working Notes), 254-262, 2019.

Other Research Papers in Related Areas

1. Gavin Abercrombie, **Aiqi Jiang**, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. Resources for Automated Identification of Online Gender-Based Violence: A Systematic Review. In The 7th Workshop on Online Abuse and Harms (WOAH), pages 170–186, Toronto, Canada. Association for Computational Linguistics, 2023.

2. Wenjie Yin*, Vibhor Agarwal*, **Aiqi Jiang***, Arkaitz Zubiaga, Nishanth Sastry. AnnoBERT: Effective Representation of Multiple Annotators' Perspectives and Label Semantics for Hate Speech Detection. In Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM 2023). (*equal contribution)

3. Xiaoyu Guo, Jing Ma, Arkaitz Zubiaga, Jianguo Xiong, Chen Zheng, **Aiqi Jiang**. A Review of Internet Meme Studies:State of the Art and Outlook. Information studies: Theroy & Application. 2021, 44(6): 199-207.

4. Arkaitz Zubiaga, **Aiqi Jiang**. Early Detection of Social Media Hoaxes at Scale. ACM Trans. Web 14, 4, Article 18 (November 2020), 23 pages, 2020.

5. **Aiqi Jiang**, Arkaitz Zubiaga. Leveraging Aspect Phrase Embeddings for Cross-Domain Review Rating Prediction. PeerJ Computer Science, 5:e225, 2019.

Declaration

This Thesis is submitted to the Queen Mary University of London in accordance with the requirements of the award of Doctor of Philosophy in the School of Electronic Engineering and Computer Science. It has not been submitted for any other degree or diploma of any examining body. Some parts of the work presented in this thesis have previously appeared in the published papers listed in the publications section. Except where specifically acknowledged, it is all the work of the Author.

Aiqi Jiang, December, 2023

Acknowledgments

Studying for a PhD is akin to exploring unknown lands, climbing sky-high mountains and crossing deep and dark valleys. My journey has been enriched by many people I have encountered along the way, and I am so grateful for their companionship and support.

First, I would like to give my sincere gratitude to my supervisor Arkaitz Zubiaga, who has always been willing and able to provide me with the support and guidance I need for the duration of this work. I also offer my big thanks to Matthew Purver and Gareth Tyson, whose feedbacks during progression vivas have been of vital importance in pursuing my PhD.

And my heartfelt thanks go to my partner Xingchen, whose patience, understanding, and unwavering support throughout the process of PhD study and this dissertation writing have been invaluable. Furthermore, I would sincerely like to thank my wonderful family members. Without them, I would not be who I am today. Their help, patience, encouragement and unconditional support throughout so many years have been fundamental in helping me reach this goal.

I also appreciate various opportunities to collaborate with researchers from Oxford Brookes University, University of Warwick, University of Surrey and Heriot-Watt University. Working alongside them has been a rewarding experience, contributing significantly to our joint research works.

I would also like to deeply thank my colleagues at the Queen Mary University of London for their role in fostering my growth as a researcher. Our regular meetings, inspiring discussions, and idea exchanges have been significant, and the leisure

moments we shared have also been cherished.

I also want to thank all the friends who have been with me all this time, not only those in my hometown in China but also my friends in the UK, for their companionship and support throughout this journey.

Finally, I would like to express my gratitude to the China Scholarship Council (CSC) for their financial support during my PhD studies. Their help has been the backbone of my academic journey.

Contents

Listing of figures

Listing of tables

Listing of acronyms

ALBERT	A Lite BERT
API	Application Programming Interface
AuxGAN	Auxiliary-loss Generative Adversarial Network
BERT	Bidirectional Encoder Representations from Transformers
BERT-wwm	BERT with whole word mask
BiLSTM	Bidirectional Long Short-Term Memory
BLI	Bilingual Lexicon Induction
CapsNet	Capsule Network
CBOW	Continuous Bag-Of-Words
CCA	Canonical Correlation Analysis
CCNL	Cross-lingual Capsule Network Learning model
CCNL-Ex	Cross-lingual Capsule Network Learning model with Extra domain-specific lexical semantics
CLSRI	Cross-Lingual Specialisation transfer based on lexical Relation Induction
CLTL	Cross-Lingual Transfer Learning

CLWEs	Cross-lingual Word Embeddings
CNN	Convolutional Neural Network
CNN-GRU	Convolutional Neural Network-Gated Recurrent Unit
DeBERTa	Decoding-enhanced BERT with Disentangled Attention
DistilBERT	Distilling BERT
DistilmBERT	Distilling multilingual BERT
ELMo	Embeddings from Language Model
et al.	et alia (en: and others)
EXIST	sEXism Identification in Social neTworks
e.g.	exemplum gratia (en: for example)
FLAN	Fine-tuned LAnguage Net
GAN	Generative Adversarial Network
GloVe	Global Vectors
GPT	Generative Pre-trained Transformer
GNN	Graph Neural Network
HAN	Hierarchical Attention Network
HPC	High Performance Computing
HurtLex	multilingual Lexicon of words to Hurt
i.e.	id est (en: that is)

JL-HL	Joint-Learning model with HurtLex
LabSE	Language Agnostic BERT Sentence Embeddings
LASER	Language Agnostic Sentence Representations
LIME	Local Interpretable Model-agnostic Explanations
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
LR	Logistic Regression
LSTM	Long Short-Term Memory
MA	MicroAggression
MacBERT	MLM as correction BERT
MAML	Model-Agnostic Meta-Learning
mBART	multilingual Bidirectional and Auto-Regressive Transformer
mBERT	multilingual Bidirectional Encoder Representations from Transformers
MLM	Masked Language Model
MLP	Multi-Layer Perceptron
mPLMs	Multilingual Pre-trained Language Models
mT5	multilingual pre-trained Text-to-Text Transfer Transformer
MuRIL	Multilingual Representations for Indian Languages
MUSE	Multilingual Unsupervised or Supervised Embedding

MWEs	Multilingual Word Embeddings
NER	Named Entity Recognition
NLP	Natural Language Processing
OOV	Out-Of-Vocabulary
PLMs	Pre-trained Language Models
POS	Part-Of-Speech
PPDB	ParaPhrase DataBase
RCSLS	Relaxed Cross-domain Similarity Local Scaling
ReLU	Rectified Linear Unit
RetroGAN	Cyclic Generative Adversarial Network-based post-specialisation system
RF	Random Forest
RoBERTa	Robustly optimised BERT approach
SA	Stereotype based on Appearance
SCB	Stereotype based on Cultural Background
SexHateLex	Sexism and Hate Lexicon
SexWEs	Sexist Word Embeddings
SHAP	SHapley Additive exPlanations
SO	Sexual Offense
SOTA	state-of-the-art

STM	Specialisation Tensor Model
SVM	Support Vector Machine
SWSR	Sina Weibo Sexism Review
TextCNN	Text-based Convolutional Neural Network
TF-IDF	Term Frequency–Inverse Document Frequency
TOCP	NTOU Chinese Profanity)
t-SNE	t-distributed Stochastic Neighbour Embedding
USE	Universal Sentences Encoder
WALS	World Atlas of Language Structures
Word2Vec	Word to Vector
XGBoost	eXtreme Gradient Boosting
XLM	Cross-lingual Language Model
XLM-R	XLM-RoBERTa
XLM-T	Cross-lingual Language Models in Twitter
XLNet	eXtreme Language understanding NETwork
XRCNN	Combination of XLM-R and TextCNN
XRCNN-Ex	Combination of XLM-R and TextCNN with External lexical knowledge

1

Introduction

Disclaimer: *Due to the nature of the research topic in this thesis, some examples may contain offensive text and hate speech. The examples do not reflect the view of the authors or their employers/graduate schools person(s), group(s), practice(s), or entity/entities. Instead they are used emphatically to help detect and prevent the spread of such harmful content.*

1.1 Motivation

The prevalence of offensive language on social media platforms, such as Facebook and Twitter, has become a pressing concern in recent years. This situation can be attributed to the anonymity provided by these platforms and the lack of strict regulations to restrain such behaviour [1]. While social media platforms have fostered connections and bridged global distances, they have unintentionally enabled the spread of hate speech and various forms of offensive language [2].

Offensive language broadly refers to expressions that may upset or annoy others, while hate speech escalates beyond mere offensive behaviour to inciting discrimination, hostility, or violence against individuals or groups based on their race, religion, ethnicity, gender, or other defining factors [3]. It specifically targets the "other" in societies, manifesting through the marginalisation of minority groups. Hate speech is one of the most important conceptual categories in anti-oppression politics today, which is considered a human rights violation in many legal systems due to its potential to cause real harm and to perpetuate discrimination against targeted groups [4, 5]. Therefore, given the severe impact of hate speech on human rights and the emergence of significant legal norms [6], there is a growing focus in the scientific community on analysing hate speech over merely offensive content. This also emphasises the need for more targeted analyses and the development of scalable and effective mitigation strategies.

Identifying hate speech is a challenging task due to its complex phenomena. One of

the primary difficulties lies in its context-dependence, where the same phrase can be benign in one situation but harmful in another [7, 8]. The context is crucial in hate speech because the interpretation of language varies widely across cultures, indicating that what is considered hate speech in one culture might not be perceived the same way in another [9, 10]. For instance, the phrase "Go back to where you came from" is often recognised as a form of racial or xenophobic hate speech directed at immigrants or people of colour in most Western countries. However, in some Asian countries, it might be interpreted as a statement about regionalism rather than a racist remark [11]. The term "bitch" is widely recognised as sexist hate speech that degrades and disrespects women. Yet, in the context of "boss bitch", it is used positively to denote a strong and independent woman, losing its derogatory connotation [12]. Another example is "chinky eyes" – a racial slur directed at East Asians. It is considered hate speech due to its derogatory and offensive nature, while "小眼睛" is often used descriptively and neutrally without any inherent racial connotations or intent to offend [13]. These examples demonstrate that various contexts can drastically alter their meaning, shifting from hateful to benign or even positive. This emphasises that effective identification of hate speech should take into account cultural variations and the specific circumstances of language use, avoiding the risk of overlooking harmful language or misinterpreting harmless expressions.

Hate speech has different types, and sexism is a common pattern of hate speech and is currently considered a deteriorating factor in social networks, especially in Asian countries [14–16]. Sex is commonly a sensitive topic, and sexist content is of high subjectivity. The high cognition and tolerance thresholds of hostile gender-biased behaviour by certain gender groups can exacerbate gender-based hatred and violence online [16]. Many existing studies focus on misogyny rather than sexism [14, 17]. However, misogyny is not always equivalent to sexism. Misogyny frequently implies the expression of hostility and hatred against women [17], while sexism is defined as an

ambivalent attitude manifested through both hostility and benevolence [18, 19]. Hostile sexism is characterised by an explicitly negative attitude towards gender groups (e.g. misogyny), while benevolent sexism is more subtle with seemingly positive characteristics. Furthermore, sexist speech refers to those promoting gender-based hate speech and violence against an individual or a gender group of people on actual or perceived aspects of personal characteristics (e.g., physical gender differences) [20], manifested in various behaviours (e.g., stereotyping, ideological issues, and sexual violence) [21, 22]. Most studies focus more on detecting hostile sexism (misogyny), overlooking implicit expressions of sexism (benevolent sexism) [14, 17]. Hence, mitigating online sexism in a wide spectrum of sexist behaviours is crucial as these are, in fact, extremely dangerous and harmful to society [23, 24].

Recognising the severity of the issue of online hate speech and sexism, the Natural Language Processing (NLP) community has proposed numerous research techniques to tackle this problem. These techniques enable the detection of offensive and sexist content based on both traditional machine learning and advanced neural network-based approaches. However, a significant linguistic gap remains, as the majority of existing studies focus only on monolingual settings (i.e., English), ignoring the multilingual nature of online offensive content. In fact, platforms such as Twitter and Facebook attract users of diverse linguistic backgrounds, and encourage them to communicate in their native languages [25], hence leading to multilingual environments. Meanwhile, existing monolingual hate speech datasets often overlook the cultural diversity within the posts and the annotators and between languages, so its subjective nature increases the difficulty of hate speech detection [10]. Due to the limited availability of labelled data and the high complexity of hate speech across diverse cultures and languages, instances of hate speech in low-resource languages are less explored [26].

Cross-Lingual Transfer Learning (CLTL) emerges as a promising solution to mitigate challenges associated with data scarcity in specific languages, by leveraging

domain knowledge from high-resource to low-resource languages [27]. At its core, it aims to utilise annotated data from diverse languages to refine detection models. This approach has the potential to improve the efficiency of offensive language detection systems, especially for languages where, due to data scarcity, development of bespoke approaches had been limited or infeasible. Given its adaptability and compatibility with neural networks, CLTL has found successful applications across various NLP tasks [28]. Some pioneering efforts have also integrated CLTL techniques for offensive language detection in low-resource languages [29–32]. However, there are still challenges in building effective and generalised cross-lingual models, particularly in understanding and bridging linguistic gaps. Cross-lingual models for hate speech detection must account for cultural variations to ensure accurate identification across diverse linguistic and cultural contexts. This involves not only translating words but also capturing the nuances, idiomatic expressions, and cultural references unique to each language, which are crucial for precise and context-sensitive hate speech detection [10, 25]. Therefore, investigating advanced CLTL methodologies for detecting online hate speech for these low-resource languages remains crucial, as it could provide new opportunities for marginalised groups as well as a deeper understanding of patterns across diverse languages and cultures.

More recently, general pre-trained language models (PLMs) have shown their capacity to improve the performance of NLP systems for most tasks on canonical data. Among the recent work for multilingual PLMs, multilingual BERT (mBERT) [33] and cross-lingual language model (XLM) [34] have stood out, thanks to the effectiveness of pre-training large transformers on multiple languages at once in the field of cross-lingual understanding [35]. However, due to the limited availability of training corpora, the XLM-RoBERTa model (XLM-R) [36] has become the new state-of-the-art (SOTA) multilingual PLMs by extending the amount of training data and the length of sentences it can handle. These SOTA PLMs are usually fine-tuned on some down-

stream classification tasks, such as multilingual and cross-lingual sexism detection [17], whereas few of them consider inducting external knowledge in a multilingual scenario into the model, such as linguistic information from a domain-specific lexicon.

Integrating structured external knowledge can yield better model performance in detecting sexist hate speech across diverse languages [37]. Domain-specific resources, such as lexicons of hateful terms and databases of reported sexist incidents, provide extra domain-related insights that enable models to differentiate between harmful speech and neutral usage. Similarly, language-specific resources, such as semantic parsers and machine translation tools, are essential for understanding linguistic diversity and cultural variations across different languages. By combining different external knowledge with the original data, we can introduce additional features or distinct lexico-semantic relations into the feature space [38, 39]. It could deepen the model's understanding of diverse sexist expressions and the linguistic variability inherent in social media, better differentiating between hateful content and neutral or contextually benign usage in various contexts [8, 40].

In this thesis, we aim to explore multilingual hate speech on social media, particularly focusing on bridging the gap in identifying sexist speech across diverse languages pertaining to different language families. Here we mainly emphasise women as the target of hate, as sexism against women is pervasive and entrenched across cultures and languages [18, 23]. To this end, we create a sexism dataset and lexicon for low-resource languages (e.g., Chinese), and leverage transfer learning techniques to develop novel models that can enable transferring the detection of sexist content from high-resource languages to low-resource languages. Additionally, we refine embedding models with domain knowledge for broader applications, thereby contributing to the identification of sexist hate speech and enhancing model understanding of discriminatory language patterns in diverse linguistic contexts.

1.2 Research Objectives

The innovative aspect of this thesis lies in the application and adaptation of transfer learning techniques to the task of hate speech detection in a cross-lingual setting. At the outset of our research in 2019, the field of cross-lingual detection of sexist hate speech was relatively underdeveloped: only five studies focused on general hate speech [40–44], with just two of these addressing sexism [40, 42]. This scarcity of research marked our research direction as not only relevant but also promising at that time. However, since then, there has been a rapid progression in the development of relevant NLP methodologies and the emergence of new datasets. This has led to some divergence between my initial works from 2019 and the latest findings from our systematic review (§3) conducted in 2023. However, these developments are natural and valuable in a dynamic field like NLP, highlighting the challenges and advances in detecting sexist hate speech across languages. My research has evolved with these changes, adapting and integrating new insights and methods wherever possible.

Therefore, the objectives of this thesis are centred around advancing the field of cross-lingual detection of online sexist hate speech, with a focus on addressing key challenges and leveraging innovative cross-lingual transfer strategies. The main objectives include:

1. **Overcoming Resource Limitations in Target Languages:** One of the main challenges in cross-lingual transfer is the lack of adequate multilingual resources, especially for low-resource languages. To address the scarcity of resources in target languages such as Chinese, we design a data collection pipeline and annotation guidelines of a sexism dataset, including the formulation of what makes a comment sexist in the context of an understudied language and culture, Chinese. We follow our designed pipeline and guidelines to create a comprehensive Chinese sexism dataset and a large domain lexicon in Chinese, as well

as release sexist word embeddings. These resources help advance research in Chinese sexism detection, and in turn the methods developed in this thesis are extensible to other low-resource languages.

2. **Bridging Discrepancies Between Source and Target Languages:** The approach should overcome linguistic and cultural discrepancies between source (high-resource languages) and target (low-resource languages) settings. This is achieved by utilising machine translation tools and alignment algorithms, thereby facilitating accurate mapping between source and target languages.

3. **Transferring Domain Knowledge at Different Hierarchies of NLP Models:** The research will focus on developing various models to achieve domain knowledge transfer at different levels. This includes instance level transfer by mapping texts between diverse languages and generating pseudo texts and labels, feature level transfer by developing sexism embedding models in the target language, as well as model level transfer by building sexism detection models for the target language or across multiple languages.

4. **Enhancing Model Understanding with External Knowledge:** To make models more domain- and language-aware, incorporating external resources can be of great benefit to improve model performance in detecting hate speech and sexism. We leverage hate speech lexicons, lexico-semantic relations, and translated data into model training.

5. **Evaluating and Refining Model Performance Across Languages:** Comparative experiments are conducted to evaluate the effectiveness of the proposed cross-lingual models against SOTA deep learning and pre-trained models.

6. **Analysing Trends in Cross-Lingual Hate Speech Detection:** Given the growing trend of analysing hate speech across languages, we conduct a compre-

hensive review of recent studies in the field of cross-lingual hate speech detection. This involves surveying the literature to analyse multilingual datasets, cross-lingual transfer scenarios and strategies, and challenges, contributing to a broader understanding of cross-lingual learning in hate speech detection.

1.3 Scope of Thesis

In this thesis, we delve into multilingual and cross-lingual hate speech detection in social media, with a specific focus on addressing sexism in cross-lingual settings. It explores the relationship and the transformation between high-resource languages (e.g., English) and other low-resource languages (e.g., Chinese). The scope of this research encompasses several key areas:

1. **Sexism Definition and Taxonomy:** We formulate the definition of sexism in the context of the Chinese language and culture, including the creation of a hierarchical taxonomy of sexism. This is developed and corrected through iterative rounds of annotations.

2. **Creation of a Chinese Sexism Dataset:** This involves collecting a dataset that includes sexist content from Chinese social media platforms (§4.2), such as Sina Weibo, and developing a comprehensive annotation guideline to label data (§4.3). A sexism-related offensive lexicon SexHateLex in Chinese is also built based on existing lexical resources (§4.4). This dataset is called Sina Weibo Sexism Review (SWSR).

3. **Cross-Lingual Model Design:** Cross-lingual models are designed by leveraging existing multilingual word embeddings (e.g., Facebook MUSE), multilingual PLMs (e.g., mBERT and XLM-R), and transfer learning architectures

(§5, §6). This also involves considering the integration of additional multilingual resources (e.g., the HurtLex hate lexicon and sentiment lexicons) (§5.2.2), and exploring the cross-lingual applicability of various NLP techniques, such as CNN/LSTM networks, pre-trained BERT model, ensemble models, and Capsule Networks (§5.3, §6.4).

4. **Refinement of Cross-Lingual Word Embeddings:** A significant part is dedicated to enhancing the semantic accuracy of pre-trained word vectors for typologically-distant language pairs (English-Chinese), specifically for sexism detection (§7). It retrofits these embeddings with external lexical knowledge and semantic specialisation techniques.

5. **Model Comparison and Analysis:** This includes a comparative analysis of the self-designed cross-lingual model against other existing models and the refined sexism-aware word embeddings against other existing embedding models in the sexism detection task (§5.3, §7.4). This comparison utilises our newly created Chinese dataset, SWSR, and other multilingual sexist benchmark datasets to assess detection performance across various languages and models.

6. **Systematic Review:** We conduct a comprehensive review of existing studies on cross-lingual hate speech (§3). This review will focus on collating insights on multilingual datasets (§3.2), cross-lingual resources (§3.3), transfer learning approaches (§3.4), and current challenges in this research area.

1.4 Structure of Thesis

This thesis consists of 8 chapters. Below we provide a brief overview summarising the contents of each of these chapters:

Chapter 1: Introduction

We present the motivation, objectives and scope for the study on identifying sexist hate speech in cross-lingual and multilingual settings, highlighting our research focus on bridging the language gap for online hate speech detection.

Chapter 2: Background and Related Work

We provide the background of hate speech and sexism, and describe the definition and workflow of hate speech detection in a cross-lingual scenario. We also present a literature review of previous works in the field, and summarise advanced techniques in related fields.

Chapter 3: Cross-lingual Transfer Learning in Hate Speech Detection

We provide a systematic overview of the recent studies on CLTL in hate speech detection. It organises and summarises all reviewed studies according to diverse aspects, including the multilingual datasets employed, cross-lingual resources leveraged, levels of knowledge transfer, and cross-lingual strategies applied. linguistic resources, and approaches utilised in CLTL. We also summarise existing challenges in the field regarding languages, datasets, and methods.

Chapter 4: Collection of Sexism Dataset and Lexicon in Chinese

We address the problem of the scarcity of Chinese resources in the field of hate speech especially for gender-related content in social media. We define a methodology for the collection and annotation of Chinese online sexism at different levels of granularity, providing the first such effort in Chinese in sexism and hate speech. A Chinese sexist lexicon is created to assist research in Chinese sexism detection. Both the dataset and lexicon are then utilised to evaluate the effectiveness of existing SOTA models in

detecting Chinese sexist content.

Chapter 5: Multi-Channel Joint Learning for Cross-Lingual Sexism Detection

We investigate the cross-lingual sexism detection task across three languages: English, Spanish and Italian, and introduce the first approach to cross-lingual sexism detection that incorporates capsule networks. We propose a cross-lingual capsule network learning model, a multi-channel architecture coupled with extra domain-specific lexical semantics (called CCNL-EX), and it yields SOTA performance for all six language pairs under study compared with ten baselines.

Chapter 6: Leveraging Pre-trained Semantics and Lexical Features for Multilingual Sexism Detection

We introduce a novel architecture based on multilingual PLMs for multilingual sexism identification using EXIST datasets, which is made of the last 4 hidden states of XLM-ROBERTA and a TEXTCNN with 3 kernels. We also exploit lexical features relying on the use of new and existing lexicons of abusive words, with a special focus on sexist slurs and abusive words targeting women.

Chapter 7: Retrofitting Sexism Domain-Aware Word Embeddings for Low-Resource Languages

We specialise the existing word embeddings with domain knowledge for one of the low-resource languages – Chinese. We develop a cross-lingual domain-aware semantic specialisation system to make the most of existing data to construct sexism-specific word embeddings, facilitating the performance of the sexism detection task for low-resource languages.

Chapter 8: Conclusion and Future Research

We summarise the main conclusions and contributions, and the outlook on the future directions of our work.

Additionally, the thesis contains the following appendices at the end:

Appendix A: Overview of Cross-lingual Hate Speech Studies

We provide two comprehensive tables that separately summarise the included dataset resources and surveyed studies in the automated identification of cross-lingual hate speech phenomena.

Appendix B: SWSR Dataset Format

We present the format and features of our Chinese sexism dataset SWSR.

2

Background and Related Work

We provide the background of hate speech and sexism, and describe the definition and workflow of hate speech detection in a cross-lingual scenario. We also present a literature review of previous works in the field, and summarise state-of-the-art (SOTA) techniques in related fields.

The chapter is organised as follows. Section 3.1 introduces the definition and difference between online hate speech and sexism, as well as available datasets and lexical resources in related fields. Section 2.2.1 describes the definition and workflow of cross-lingual hate speech detection. In Sections 2.2.2 and 2.2.3, we summarise the previous efforts towards detecting online hate speech and sexism respectively in the cross-lingual scenario. Then we summarise the recent works in which cross-lingual detection has been profited to external domain knowledge integration in Section 2.3.

2.1 Hate Speech and Sexism in Social Media

2.1.1 Background

Along with an unprecedented ability for communication and information sharing, social media platforms provide an anonymous environment which allows users to take aggressive attitudes towards specific groups or individuals by posting abusive language. This leads to increased occurrences of incidents, hostile behaviours, and remarks of harassment [45–48]. Abusive language is one of the most important conceptual categories in anti-oppression politics today [4, 45]. An example of abusive content is the posting of hate speech, i.e. the use of language to incite violence, promote hatred or disparage individuals or communities on the basis of specific characteristics, such as religion, gender, ethnicity or sexual orientation [1, 49, 50].

Sexism is a common pattern of hate speech and is currently considered a deteriorating factor in social networks [14–16]. Sex is a sensitive topic in Asian cultures, hence many women still have a high cognitive and tolerance threshold for hostile gender-

biased behaviours [16], which consequently aggravates abusive remarks and violent behaviours online. The task of mitigating hate speech online has attracted the attention of industries to impose strict censorship on the contents of relevant topics [51], but has remained largely understudied in academic research [52].

In the past few years, due to the increasing amount of user-generated content and the diversity of user behaviour towards women in social media, manual inspection and moderation of sexist content has become unmanageable. The academic community has seen a rapid increase in research tackling the automatic detection of misogynous behaviour and gender-based hatred in both monolingual and multilingual scenarios [53, 54].

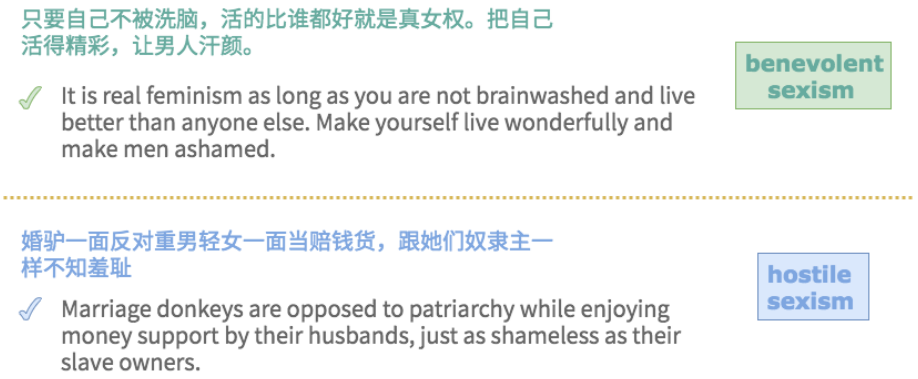

Figure 2.1: Examples of hostile and benevolent sexism.

However, misogyny is not always equivalent to sexism, and frequently implies the expression of hostility and hatred against women [17]. As for sexism, Glick and Fiske [18] define the concept of sexism referring to two forms of sexism: hostile sexism and benevolent sexism. Hostile sexism is characterised by an explicitly negative attitude towards women, while benevolent sexism is more subtle with seemingly positive characteristics (see examples in Figure 2.1). Sexism includes a wide range of behaviours (such as stereotyping, ideological issues, sexual violence, etc.) [21, 22], and may be expressed in different ways: direct, indirect, descriptive or reported [55, 56]. Thus,

misogyny is only one case of sexism [21]. Most previous studies concentrate more on detecting hostile and explicit sexism, overlooking subtle or implicit expressions of sexism [14, 15, 17, 22]. Hence, dealing with the detection of sexism in a wide spectrum of sexist attitudes and behaviours is necessary as these are, in fact, the most frequent and dangerous for society [23].

2.1.2 Sexism Datasets

Most relevant studies for identifying online abusive content against women utilise supervised approaches, and recently, deep learning approaches have become more popular, especially transformer-based approaches, which have made SOTA achievements in different languages [54, 56–58]. Since these approaches for automatic sexism detection are usually established utilising labelled training data, the performance is more dependent on the quality and taxonomy of the available datasets [59]. The last few years have witnessed an increase in the interest in and availability of sexism datasets. The earliest attempt was by Waseem and Hovy [14], who provided a publicly available dataset of more than 16k tweets for hate speech and annotated it into three categories – racism, sexism and neither. However, it only comprises the expression of hostile sexism towards women, overlooking other kinds of sexism. Chowdhury et al. [60] aggregate experiences of sexual abuse to facilitate a better understanding of social media construction and to bring about social change. These two datasets consist of content in English.

In addition, recent sexism datasets include multilingual content involving Italian, Spanish and Hindi, along with English. The Automatic Misogyny Identification (AMI) competitions in Evalita 2018 [46], Ibereval 2018 [61] and Evalita 2020 [47] provide datasets in English, Spanish and Italian to detect misogynistic content, to classify misogynous behaviour as well as to identify the target of a misogynous text. HatEval@SemEval 2019 [62] is another competition aiming to detect hate speech

against immigrants and women and further finer-grained features in offensive text, like aggressive attitude and the target harassed in English and Spanish posts from Twitter. Furthermore, Parikh et al. [57] introduce a dataset consisting of accounts of sexism in 23 categories to investigate sexism categorisation as a multi-label classification task. Bhattacharya et al. [63] develop a multilingual annotated corpus of misogyny and aggression in Indian English, Hindi, and Indian Bangla as part of a project studying and automatically identifying misogyny and communalism in social media. The first French dataset [48] and Spanish dataset (MeTwo) [54] have been released for sexism detection, and EXIST@IberLEF 2021[1] proposes the first shared task on sexism identification in social networks (as opposed to misogyny detection), aiming to detect online sexism in English and Spanish. Moreover, Mulki and Ghanem [64] introduce the first Arabic Levantine dataset for online Misogyny (LeT-Mi) written in the Arabic and Levantine dialect. Then ArMI@HASOC 2021 at FIRE[2] proposes an Arabic Misogyny Identification (ArMI) task with two sub-tasks derived from the Let-Mi dataset [64], which is the first shared task to address the problem of automatic detection of Arabic online misogyny. Guest et al. [65] introduce an expert annotated misogynous dataset collected from Reddit and present a new detailed hierarchical taxonomy for online misogyny, while Zeinert et al. develop the first Danish misogyny dataset, Bajer, under a four-level taxonomy of labels [66]. Besides, Samory et al. [58] provide a sexism dataset using psychological scales and generating adversarial samples to improve construct validity and reliability in sexism detection.

Most popular social media, such as Twitter and Facebook, are highly multilingual. They foster their users to interact in their primary language. So there is a considerable urgency to develop a robust approach to sexism detection in a multilingual environment. However, most existing research on sexism detection focuses on the English language [14], due to its advantageous position over other languages in terms of

[1] http://nlp.uned.es/exist2021
[2] https://sites.google.com/view/armi2021/

available resources. This in turn leads to a dearth of research in other languages [1]. The recent trend is increasingly focusing on investigating sexism detection in Indo-European languages [46, 61, 63, 64]. The dataset creation for low-resource languages poses several challenges when it comes to data collection and annotation, especially with the diversity of dialects and the ambiguity brought about by the emerging Internet languages.

2.1.3 Lexical Resources for Online Abuse

Detection of offensive content can be challenging as it does not always contain explicit mentions of negative or hateful words [67, 68]. However, there is evidence showing that the use of domain-specific lexical words in classification models can boost model performance [37, 69, 70]. With the expectation that the use of a lexicon can make for a good proxy to improve the detection of hate speech, in this thesis we develop one in Chinese to support our research in sexism detection. There are many popular lexicons for online abuse, which collect and organise offensive words and phrases. For example, Burnap and Williams [71] focus on several lists obtained from Wikipedia that are particularly linked to a specific sub-type of hate speech in English, such as ethnic slurs[3] and LGBT slang terms.[4] A popular hate speech lexicon is HateBase,[5] which provides the largest multilingual hate speech lexicon linked to aspects such as religion, gender and ethnicity. It includes 3,635 groups of terms in more than 95 languages [72]. Despite its volume for languages like English, the HateBase lexicon only contains 39 Chinese terms, which lacks the potential to be effective at scale and is still far from becoming a referential resource. Besides, Bassignana et al. [73] built a multilingual hate speech lexicon, HurtLex,[6] involving over 50 languages. HurtLex

[3] https://en.wikipedia.org/wiki/List_of_ethnic_slurs
[4] https://en.wikipedia.org/wiki/List_of_LGBT_slang_terms
[5] https://hatebase.org/
[6] https://github.com/valeriobasile/hurtlex

is one of the resources that we leverage in this thesis to build a new Chinese lexicon.

2.2 Cross-lingual Transfer Learning

2.2.1 Task Description and Workflow

The task of cross-lingual hate speech detection focuses on the identification and categorisation of offensive or hateful text across different languages. The primary objective for CLTL is to leverage the knowledge acquired from one language (source language with more resources) to enhance the hate speech detection in another language (target language with less resources), especially when the labelled data in the target language is scarce or unavailable [25]. The workflow of cross-lingual hate speech is presented below:

Data Preparation

Let L_s and L_t represent the source and target languages respectively. Then let $D_s = \{X_s, Y_s\}$ and $D_t = \{X_t, Y_t\}$ represent the source and target datasets respectively, where X denotes text data and Y denotes labels indicating that the text is offensive or not. The cross-lingual task $CL = \{Y_t, f : F \rightarrow Y_t\}$ contains a predictive function f, a feature space F, and a label space $Y_t = \{0, 1\}$, where 0 denotes non-offensive and 1 denotes offensive content.

Cross-lingual Training

The training stage for cross-lingual transfer can use only D_s or both D_s and D_t. For only D_s as training data, a model M_s is trained to learn a function $f_s : F_s \rightarrow Y$, where F_s is the feature space of L_S. Then adapt the learned function f_s to a function $f_t : F_t \rightarrow Y$, where F_t is the feature space of L_t, with minimal loss in performance. For both D_s and D_t as training data, a model M_{st} is trained to learn a function

$f_{st} : F_{st} \rightarrow Y$, where F_{st} is the joint feature space of L_{st}. Various cross-lingual training strategies could be employed for cross-lingual transfer to obtain f_t, where more details are described in Section 3.4.

Model Adaptation

If label Y_t is available for D_t, fine-tune f_t on full or few shots of D_t to obtain a model M_t, further adapting the function to the target domain: $f_t = finetune(f_t, D_t)$.

Detection on Target Language

Apply predictive functions f_t or f_{st} to D_t to detect offensive content in L_t and predict labels: $\hat{y}_t = f_t(D_t)$ or $\hat{y}_t = f_{st}(D_t)$.

2.2.2 Cross-lingual Studies in Hate Speech Detection

With the prevalence of online social media, a range of Natural Language Processing (NLP) approaches, especially transformer-based techniques [35], have been employed to identify online hate speech [1, 74, 75] or focusing on detecting specific types of hate, such as racism [76, 77], sexism [14, 17], and cyberbullying [71, 78], but limited to a single language, generally in English.

While more mature areas of NLP such as sentiment analysis have accumulated substantial efforts by employing cross-lingual learning techniques, work on hate speech detection in cross-lingual scenarios has not been explored as much [79]. It does however bring additional challenges compared to the sentiment analysis, as how hate is expressed across different languages and cultures varies. Basile and Rubagotti [80] use SVM with n-grams to tackle English and Italian in a cross-lingual setting in Evalita 2018 and achieve the 15th/2nd position for English/Italian. Pamungkas and Patti [40] propose a joint-learning cross-lingual model with multilingual HurtLex [73] and MUSE embeddings [81], which outperforms other models using monolingual embeddings [40].

Several multilingual multi-aspect approaches are conducted for hate speech [82] and cross-lingual contextual word embeddings are applied in offensive language identification from English to other languages [83, 84]. Due to the scarcity of cross-lingual resources in this field, some studies tend to generate parallel corpora directly leveraging machine translation resources such as Google Translate [17, 40, 85, 86], which has proved the effectiveness of the approach.

2.2.3 Cross-lingual Studies in Sexism Detection

Research in social media sexism detection has increased in recent years [57, 58]. The first attempt was by Hewitt et al. [87] who investigated the manual classification of gender-based tweets, and the first survey of automatic misogyny identification in social media was conducted by Anzovino et al. [22]. Rodríguez-Sánchez et al. [54] explore the feasibility of automatically identifying sexist content using both traditional machine learning and deep learning techniques.

In addition, researchers mainly address the problem of multilingual sexism detection by using deep neural networks with cross-lingual word embeddings (CLWEs) or multilingual PLMs [17, 88]. However, most relevant studies investigate monolingual or multilingual sexism detection only based on existing data in high-resource languages such as English and other Indo-European languages [47, 48, 58], while cross-lingual studies in the field of sexism and even general abuse are still limited for low-resource languages like Chinese.

2.3 Integrating External Features

Most studies in sexism detection focus on investigating the superior model architecture for classification in different languages (e.g., neural network-based and transformer-based architectures) [54, 57, 58], and few make efforts to infuse external domain knowledge into the vector space to enhance the detection performance [89]. Several

works demonstrate the positive influence on the broader abusive language detection task by directly injecting external domain knowledge at the model level [37], but only a few perform an exploration into the effect of this knowledge. Badjatiya et al. [50] utilise an LSTM-based model to generate English hate word embeddings, but more persuasive validation strategies should be reconsidered [74]. Kamble and Joshi [90] describe the construction of domain word embeddings based on Word2Vec from a Hindi-English hate speech dataset, and Alatawi et al. [91] produce abuse-specific embeddings for English white supremacy. Besides, multilingual word embeddings based on abuse knowledge are created for cross-lingual hate speech detection [89].

Incorporating additional features to specialise word vectors could be of benefit, and there has been a body of research exploring various methods to incorporate diverse constraints into the word embedding space. The first retrofitting work by Faruqui et al. [38] is proposed to pull the vectors of similar words closer to each other by fusing only synonyms. Then ATTRACT-REPEL, a standard semantic specialisation approach, is developed to integrate structured linguistic constraints with both similar and dissimilar semantics into pre-trained vector spaces, clustering the embeddings of similar words (e.g., synonyms, hypernym-hyponym pairs) closer together and enforcing dissimilar words (e.g., antonyms) far away from each other [39]. Such semantic specialisation could be applied to any kind of distributional word embeddings.

Since the first-generation semantic specialisation models only retrofit the embeddings of words seen in linguistic constraints, a series of post-specialisation techniques are proposed [92–95]. Post-specialisation aims to fine-tune the entire distributional vector space by learning an explicit and global specialisation mapping between original and initially specialised spaces, and then applying the mapping to the embeddings of words unseen in external constraints [92]. Ponti et al. [94] and Colon-Hernandez et al. [95] modify the feed-forward post-specialisation network with different Generative Adversarial Networks (GAN) based approaches to discriminate word vectors from

original and specialised spaces, which yields better performance on retrofitting.

Post-specialisation approaches can be further employed for cross-lingual transfer through a shared vector space between source and target languages [96, 97]. In this thesis, we demonstrate how to combine task-oriented multilingual domain knowledge to achieve cross-lingual semantic specialisation on pre-trained word embeddings, with an impact on sexism detection for low-resource languages.

3

Cross-lingual Transfer Learning in Hate Speech Detection

In this chapter, we provide a systematic overview on the literature of cross-lingual transfer learning in hate speech detection. It describes all reviewed studies according to diverse aspects, including the multilingual datasets employed, cross-lingual resources leveraged, levels of knowledge transfer, and cross-lingual strategies applied. Challenges are also summarised in three aspects: language, dataset and approach.

The chapter is organised as follows. Section 3.1 introduces our survey work and distinguishes "cross-lingual" from related terminologies such as "multilingual" and "code-mixing". Then we summarise and analyse the multilingual datasets leveraged in surveyed papers in Section 3.2. In Section 3.3, we describe diverse linguistic resources and tools that are often used in cross-lingual studies. Section 3.4 reveals the three transfer levels of CLTL in offensive language detection, complemented by transfer strategies based on these layers. Section 3.5 presents current challenges in this area.

3.1 Introduction

Cross-Lingual Transfer Learning (CLTL) emerges as a promising solution to mitigate challenges associated with data scarcity in specific languages, by leveraging domain knowledge from high-resource to low-resource languages [27]. The earliest work is from 2018 [44] and since then the interest in cross-lingual offensive language detection has increased. Although the number of research in this area has increased significantly, however to the best of our knowledge, comprehensive surveys remain elusive. Existing surveys often concentrate on monolingual detection solutions or specific characteristics of offensive language, while the unique CLTL application in offensive language detection is seldom considered. A couple of relevant surveys do exist [2, 25], but their scope goes beyond cross-lingual scenarios, also covering cross-domain and multimodal dimensions. Our focus remains on the application of CLTL to offensive language detection.

Our objective is to systematically review existing literature, elucidating different

CLTL techniques employed in the task of offensive language detection. Consequently, this chapter discusses 67 papers surveyed in this area, and the distribution of their publication year is shown in Figure 3.1. This is the first systematic and holistic overview of recent studies using CLTL to detect offensive language in social media across languages. We describe those studies according to diverse aspects, including the multilingual datasets employed, cross-lingual resources leveraged, levels of transfer, and cross-lingual strategies applied. Inspired by Pikuliak et al. [27], we outline three CLTL transfer approaches based on the level at which knowledge transfer occurs across languages, namely instance transfer, feature transfer, and parameter transfer. We also illustrate prevalent transfer strategies utilised based on these transfer levels. In addition, we highlight current challenges and aim to provide several potential opportunities for future research in this field. Furthermore, we present two comprehensive tables in Appendix A containing multilingual datasets and CLTL techniques used in surveyed papers respectively to facilitate easy comparison and discovery of related works.

Cross-lingual & Multilingual & Code-mixing

Cross-lingual and multilingual learning approaches both delve into challenges across multiple languages. While "cross-lingual learning" typically denotes the application of CLTL techniques, the term "multilingual learning" is sometimes used interchangeably with it [27]. Some multilingual data may contain code-mixing content, where users express their opinions using a mixture of languages in the same sentence [43]. Hence multilingual learning includes the special case of code-mixing. Although some researchers draw parallels between these terms and the relationship between transfer learning and multi-task learning [98], we adopt a more expansive perspective on cross-lingual learning in our survey. Essentially, we define cross-lingual learning as the process of transferring knowledge between different languages, a concept that in-

herently encompasses multilingual learning. Thus, we regard multilingual learning as a subset of cross-lingual learning.

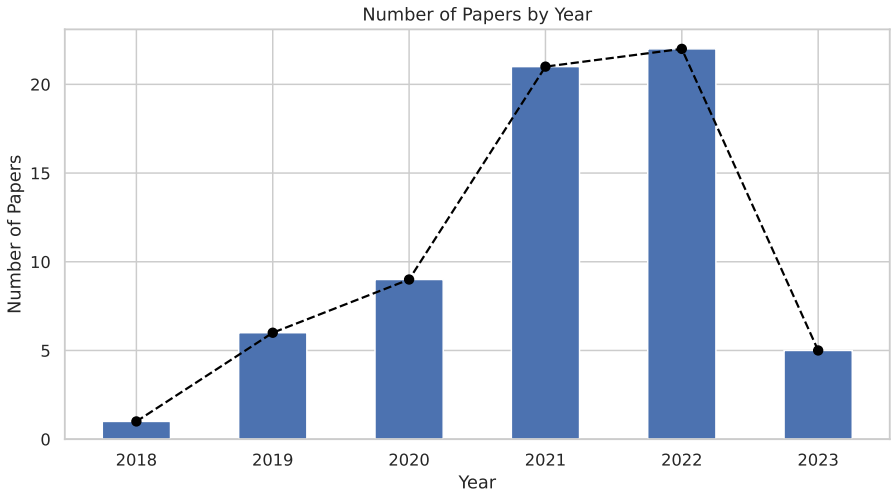

Figure 3.1: Publications per year up to July 2023.

3.2 Multilingual Hate Speech Datasets

Multilingual datasets provide the necessary data across multiple languages to facilitate cross-lingual transfer of hate speech knowledge, bridging the language gap and mitigating hateful content in multilingual scenarios. The foundation of robust machine learning models is the quality and comprehensiveness of the datasets they are trained on, especially for supervised models. More diverse datasets could offer opportunities to alleviate the data scarcity problem, enhance cross-lingual model generalisability, and facilitate comparative studies. We investigate 82 datasets utilised in existing surveyed studies on cross-lingual hate speech detection across different aspects. We summarise these datasets into Table A.1 in terms of their publication year, topic of hate speech, data source for collection, languages, number of instances, percentage of hateful texts, type of labels, number of annotators for data labelling,

number of citations, whether or not the dataset is code-mixed, and whether or not the dataset is available.

Topics

These datasets cover a variety of topics and employ different terms to describe the types of offensive content. The topic distribution among datasets shows that "offence" and "hate speech" are the two most frequently addressed topics, accounting for 34.23% and 32.43% respectively. There are also other topics focusing on general offence, such as abusiveness, toxicity, cyberbullying, harassment and aggressiveness. In addition, some datasets narrow their scope to specific issues, like sexism, misogyny and racism, targeting particular individuals or groups. When detecting cross-lingual hate speech, inherent biases between topics within these datasets may arise. The inconsistency in terminology, data collection and annotation can introduce potential biases [99, 100].

Data Sources

While hate speech is prevalent in all online spaces, most studies tend to collect data from freely accessible social media sources such as Twitter and Facebook. Our review of data sources reveals a pronounced reliance on popular social media platforms, with Twitter notably constituting 47.19% of the datasets. This is followed by YouTube and Facebook, which contribute to 11.24% and 10.11% respectively. The dominance of these platforms in the data collection underscores their ubiquity and the pressing need to monitor and mitigate hate speech on such widely-used platforms. While these mainstream platforms dominate, news websites like Fox News and NAVER news, and open forums like Reddit and 2ch, also contribute to the multilingual datasets. And there is a rich diversity in the data source origins. Sources like Weibo (Chinese), 2ch (Russian), g1.globo (Brazilian) and Eesti Ekspress (Estonian) represent some non-English speaking regions, which can be more culturally specific platforms. Some

datasets originate from platforms known for controversial content or specific user bases, such as Stormfront and alt-right websites. These sources, though less frequent, emphasise the pervasiveness and more extreme forms of hate speech across different online media.

Languages and Families

All datasets cover 32 distinct languages among 10 language families, where one dataset may contain more than one language or family. We show the distribution of languages and their language families in Figures 3.2 and 3.3 respectively. The distribution of languages in the datasets highlights a representative language of English, accounting for 25.95% of the data, followed by Hindi at 9.16%, and German at 6.87%. Notably, the datasets exhibit a relatively balanced representation across several languages, such as Spanish, Arabic, Italian, French, Portuguese and Turkish, each ranging from around 2% to 5%. This distribution reveals a strong emphasis on Indo-European languages in hate speech studies, complemented by a notable presence of Afro-Asiatic languages (represented primarily by Arabic), with other language families being less represented and restricting cross-lingual research. It is also noteworthy that some datasets (14.6%) are dedicated to code-mixed content, a mixture of two or more languages. This is a common phenomenon in non-English and multilingual societies and online platforms, adding to the complexity of real-world hate speech use.

Size

The distribution of dataset size is illustrated in Figure 3.4. Over half of the datasets, accounting for 51.20%, contain in the range of 10^4-10^5 instances. This is closely followed by 35.40% of datasets that have instances between 10^3 and 10^4. On the other hand, large datasets are rare, with only one dataset's size exceeding 10^7, and two between 10^6 and 10^7. We can observe that the majority of datasets fall within the

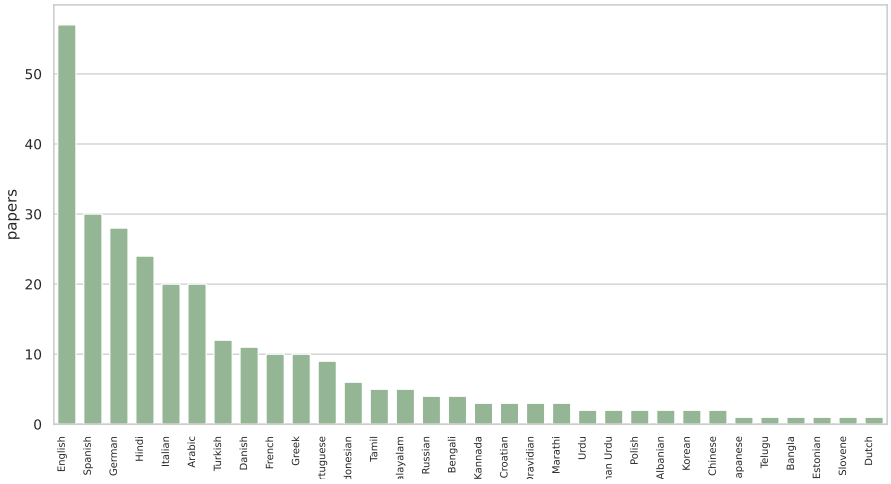

Figure 3.2: Distribution of languages covered by the datasets used in the surveyed studies of cross-lingual hate speech detection.

smaller size ranges, which indicates potential challenges in collecting and manually annotating large-scale labelled datasets, as well as model generalisability for cross-lingual hate speech research.

Data Labelling and Distributions

According to diverse annotation schemes, four different types of labels are often employed: (i) binary labels, (ii) fine-grained categories of offensive content, (iii) attack targets, and (iv) intensity scores. A majority of datasets, 76 in total, utilise straightforward binary labelling, whereas 36 datasets only provide binary labels. There are also datasets combining binary labels with either fine-grained categories (20), attack targets (9), or intensity scores (1). Besides, a smaller subset of datasets (only 10 in total) include binary labels, fine-grained categories, and attack targets, offering a more comprehensive annotation. Interestingly, only four datasets consider intensity scores, suggesting that quantifying the severity of offensive content is less common in current research.

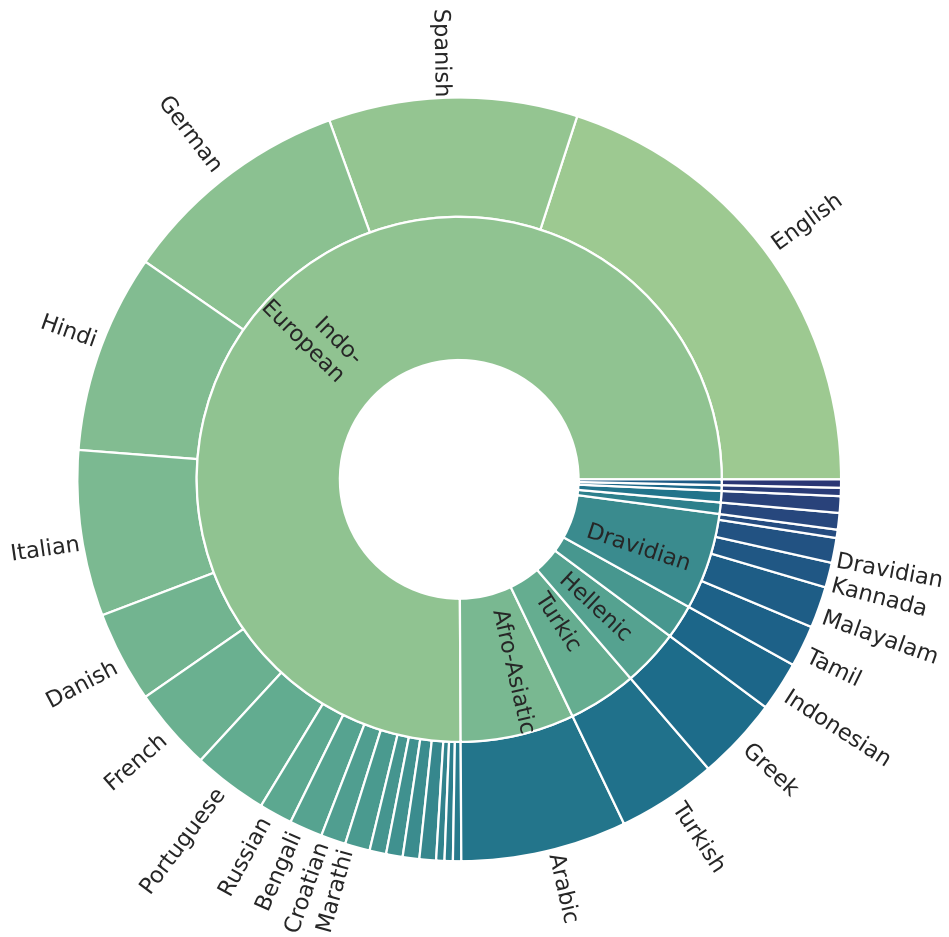

Figure 3.3: Distribution of languages and language families covered by the datasets used in the surveyed studies of cross-lingual hate speech detection. The inner pie chart shows the proportion of each language family, and the outer pie chart shows the proportion of each language corresponding to its family.

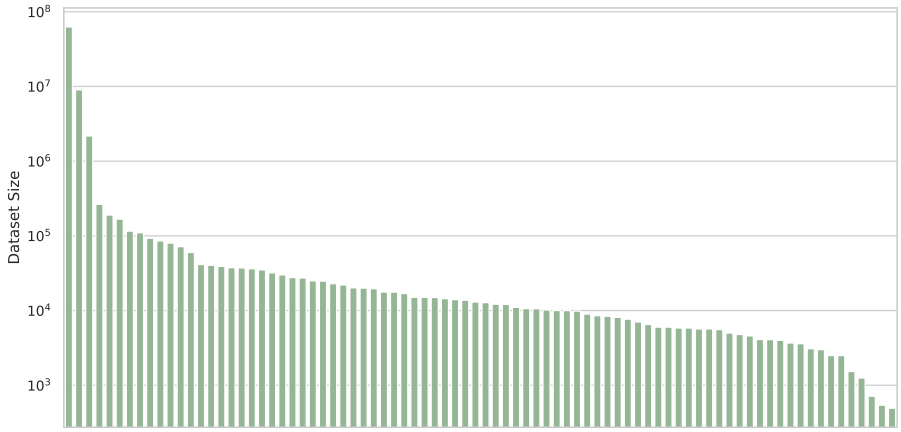

Figure 3.4: Distribution of dataset sizes used in surveyed studies of cross-lingual hate speech detection.

Dataset Availability

Across the surveyed datasets, we find 66 out of 82 data resources are readily accessible to researchers, fostering transparency and reproducibility in this field. However, 16 resources remain unavailable, possibly due to proprietary constraints and privacy considerations on certain platforms, or exclusivity to specific research groups or institutions. For some resources, researchers must directly contact the authors to request access.

Competitions and Shared Tasks

The rising concern surrounding online hate and related phenomena has led to the establishment of numerous competitions and shared tasks within both national (e.g., GermEval, Evalita, IberLEF) and international (e.g., SemEval) evaluation campaigns. These open scientific competitions release benchmark datasets and invite participants to submit detection results and detailed system reports [101]. Table 3.1 lists 24 such

competitions and shared tasks from 2018 to 2023, each addressing different facets of offensive content across multiple languages.

Prominent topics include offence and hate speech. Specifically, GermEval competitions, conducted in 2018 [102] and 2019 [103], centre on identifying offensive content in German tweets. OffensEval, initiated in 2019 for only English content [104], expands to include multiple languages in the subsequent edition in 2020 [105]. The task in OSACT4 [106], the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, delves into the challenges of offensive language detection in Arabic, especially its dialectal forms. HaSpeeDe has consistently addressed hate speech, with its iterations focusing on various general and particular hate speech topics [107–109]. For instance, HaSpeeDe [107] and HaSpeeDe2 [108] concentrate on hate speech against immigrants and Muslims, while HaSpeeDe3 [109] explores hate speech in strongly polarised debates, in particular concerning political and religious topics. Similarly, the HASOC series, recurring annually since 2019, delves into hate speech and offensive content identification across various languages [110–113]. TRAC workshops, held in 2018 [114] and 2020 [115], spotlight trolling, aggression, and cyberbullying. Other competitions, like PolEval [116] and UrduThreat [117], focus on specific languages (such as Polish and Urdu) or topics (e.g., cyberbullying and threatening languages). Kaggle is a platform for hosting predictive modelling and analytics competitions among the global community of data scientists. The Jigsaw Multilingual Toxic Comment Classification competition on Kaggle witnesses participation from over 1600 teams, encouraging them to build multilingual models using English-only training data to identify diverse toxic content. Another Kaggle competition, IIIT-D Multilingual Abusive Comment Identification, focuses on abusive comments in Indic languages. Additionally, some tasks focus on more specific topics, such as sexism and misogyny. AMI is organised by IberEval [61] and Evalita [46] in 2018, specifically addressing Spanish and Italian misogynous content respectively. EXIST is held for three consecutive years

from 2021 to 2023 and aims to identify sexism in social networks [88, 118, 119].

The majority of these benchmark datasets originate from social media platforms like Twitter, and often conduct annotations that go beyond simple binary labels to finer-grained ones. Their online availability accelerates research in cross-lingual hate speech detection. Besides, the substantial participation in these shared tasks, within a relatively short timeframe, not only underscores the global community's interest in hate speech detection but also motivates the continuation of such competitions and shared tasks.

3.3 Cross-lingual Resources

3.3.1 Multilingual linguistic Resources

Some basic but essential linguistic resources facilitate the task of cross-lingual hate speech detection across two or more languages. Among these, multilingual lexicons and parallel corpora stand out as foundational resources, frequently utilised by researchers to bridge linguistic gaps and enhance model performance in CLTL.

- *Multilingual Lexicons (word-aligned):* contain words, phrases, or terms in two or more languages. These lexicons often provide direct translations or equivalents of terms across the languages they cover, such as HurtLex.[7]
- *Parallel Corpora (sentence-aligned):* refer to datasets that consist of texts in two or more languages, where each text in one language has a direct translation in the other languages.

Due to the lack of sufficient labelled datasets for hate speech, multilingual lexicons can help bridge this gap by providing connections between resource-rich and low-resource languages [29, 120], while parallel corpora provide translations of labelled

[7] https://github.com/valeriobasile/hurtlex

Name	Description	Year	Topic	Language	#Teams
GermEval [102]	Identification of Offensive Language	2018	offence	de	20
AMI@IberEval [61]	Automatic Misogyny Identification	2018	misogyny	en, es	11
AMI@Evalita [46]	Automatic Misogyny Identification	2018	misogyny	en, it	16
HaSpeeDe@Evalita [107]	Hate Speech Detection	2018	hate speech	it	9
TRAC-1 [114]	Aggression Identification of Hindi-English Code-mixed Data	2018	aggressiveness	en, hi	30
HatEval@SemEval [62]	Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter	2019	hate speech	en, es	74
HASOC@FIRE [110]	Hate Speech and Offensive Content Identification in Indo-European Languages	2019	hate speech, offence	en, hi, de	37
Task2@GermEval [103]	Identification of Offensive Language	2019	offence	de	13
Task6@PolEval [116]	Automatic Cyberbullying Detection in Polish Twitter	2019	cyberbullying	pl	9
OffensEval@SemEval [105]	Multilingual Offensive Language Identification in Social Media	2020	offence	en, ar, da, el, tr	145
HASOC@FIRE [111]	Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German	2020	hate speech, offence	en, hi, de, ta, ml	40+
HaSpeeDe@Evalita [108]	Hate Speech Detection	2020	hate speech	it	14
TRAC-2 [115]	Aggression and Gendered Aggression Identification	2020	aggressiveness	en, hi, bn	19
Jigsaw Toxic@Kaggle	Jigsaw Multilingual Toxic Comment Classification	2020	toxicity	en, es, tr, pt	1621
OSACT4 [106]	Arabic Offensive Language Detection	2020	offence	ar	27
HASOC@FIRE [112]	Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech	2021	hate speech, offence	en, hi, mr	65
EXIST@IberLEF [88]	Sexism Identification in Social Networks	2021	sexism	en, es	31
UrduThreat@FIRE [117]	Abusive and Threatening Language Detection in Urdu	2021	abuse, threat	ur	19
IIIT-D@Kaggle	Moj Multilingual Abusive Comment Identification across Indic Languages	2021	abuse	10+ Indic	54
HASOC@FIRE [113]	Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages	2022	hate speech, offence	en, hi, de, mr	12
EXIST@IberLEF [118]	Sexism Identification in Social Networks	2022	sexism	en, es	19
HaSpeeDe@Evalita [109]	Political and Religious Hate Speech Detection	2023	hate speech	it	6
EXIST@IberLEF [119]	Sexism Identification in Social Networks	2023	sexism	en, es	28
HASOC@FIRE	Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages	2023	hate speech, offence	en, hi, de, Indo-Aryan	-

Table 3.1: Summary of competitions and shared tasks in automated identification of cross-lingual hate speech. Relevant monolingual tasks in non-English languages are also included. Language names are represented by using the standardized nomenclature ISO 639-1. "#Teams" = "number of teams participated in the competition and submitted runs". All task links are added to the "Name" column.

instances from resource-rich to low-resource languages, to facilitate CLTL [121–123]. CLWEs are also created by utilising these multilingual resources. In addition, these resources are also valuable linguistic resources in other Natural Language Processing (NLP) tasks, such as machine translation and multilingual model training.

Multilingual linguistic resources are most often created for specific domains, such

as hate speech and abusive language. While certain hate speech patterns can be universal, others can be specific to certain cultures or languages. These resources provide insights into how hate speech manifests differently across languages and cultures, furthering the understanding of cultural nuances [79].

3.3.2 Machine Translation and Transliteration Tools

Translation refers to converting content from one language into another while preserving the meaning of the original text, which is a complex task involving a deep understanding of the full linguistic context. Transliteration, on the other hand, is the process of converting text from one script into another, while preserving only the phonetic aspects without changing the actual meaning [124]. Machine translation tools can translate datasets from one language to another, such as Google Translate,[8] Microsoft Translator,[9] DeepL,[10]) and machine translation models (e.g., mBART [125] and mT5 [126]). Besides, machine transliteration tools are essential for code-mixed datasets, such as Google Transliteration,[11] Microsoft Transliteration,[12] AI4Bharat transliteration application,[13] and transliteration python packages (e.g., indic-transliteration[14]).

However, some translation tools may inadvertently bring out translation errors or cultural nuances that are inconsistent with the original meaning [120]. For instance, certain derogatory terms or slurs might not have direct equivalents in other languages, or their severity might differ across cultures. Additionally, idiomatic expressions that convey hate or abuse in one language might lose their offensive meaning when translated literally. Therefore, the quality of translated datasets is crucial to maintain

[8] https://translate.google.com
[9] https://translator.microsoft.com/
[10] https://www.deepl.com/en/translator
[11] https://www.google.co.in/inputtools/try/
[12] https://www.microsoft.com/en-us/translator/business/translator-api/
[13] https://github.com/AI4Bharat/IndianNLP-Transliteration
[14] https://github.com/indic-transliteration/indic_transliteration_py

the effectiveness and robustness of cross-lingual detection models.

3.3.3 Multilingual Representations

Multilingual representations, as language-independent representations, are often used in CLTL. With these representations, we can directly address the difference between source and target languages by projecting the text into a shared feature space. Then it is easier to project the behavior of the model. In general, we can distinguish between using word-level representations and sentence-level representations.

Multilingual distributional representations represent words from multiple languages in a single distributional word vector space. Cross-lingual transfer of word embeddings aims to establish the semantic mappings among words in different languages by learning the transformation functions over the corresponding word embedding spaces. In this space, semantically similar words are close together independently of the language they come from. For example, cat in English and katze in German should have geometrically similar vector representations. Some prominent distributional embeddings are Multilingual GloVe [127], Multilingual FASTTEXT [128], Babylon [129], and Multilingual Unsupervised or Supervised word Embeddings (MUSE) [81, 130]. During the training, multilingual distributional representations usually require additional cross-lingual resources, e.g., bilingual dictionaries or parallel corpora.

Multilingual contextualised representations on sentence level work on a similar principle, but they use sentences instead of words for alignments. They are usually based on an auto-encoder architecture, in which the model is pushed to create similar and context-aware representations for parallel sentences, such as Language Agnostic Sentence Representations (LASER) [131] and Language Agnostic BERT Sentence Embeddings (LabSE) [132].

3.3.4 Monolingual and Multilingual Pre-trained Language Models

Pre-trained language models (PLMs) are a state-of-the-art (SOTA) technique in NLP, which is trained by a large amount of data. PLMs can be initialised and further trained or fine-tuned with target data for CLTL. The most well-known examples of PLMs are Embeddings from Language Models (ELMO) [133] and Bidirectional Encoder Representations from Transformers (BERT) [33].

Multilingual pre-trained language models (PLMs) are an extension of monolingual PLMs, trained with multiple languages at the same time. They learn to understand the connections between languages by developing a shared linguistic representation. There is no need to provide multilingual PLMs with any additional information about interlanguage relations, and cross-lingual transfer is possible between any languages [27]. Prominent examples include Multilingual BERT (mBERT) [33] and XLM-RoBERTa (XLM-R) [36], which have demonstrated remarkable capabilities in understanding and processing multiple languages simultaneously.

3.3.5 Language-Agnostic Resources

Language-agnostic resources (e.g., emojis, capitals, and punctuations) can be seen as common traits among different languages, sharing similar knowledge to enhance the linguistic connection in cross-lingual learning. For instance, emojis are associated with emotional expressions, which in turn are associated with various forms of online hate [134]. They often have high coverage in social media texts.

3.4 Cross-lingual Transfer Approaches

In this section, we systematically describe and compare different cross-lingual techniques to detect hate speech content online. Referring to the way of identifying different transfer learning techniques in [27] and [28], we analyse cross-lingual approaches

in all reviewed papers and ultimately decide to categorise them according to *"what to transfer"* – the level at which knowledge transfer or sharing occurs.

The overall hierarchy of various cross-lingual transfer approaches is depicted in Figure 3.5. Our cross-lingual transfer paradigms consist of three main categories: instance transfer, feature transfer, and parameter transfer.

- **Instance Transfer:** Instances are transferred on the data level between source and target languages, including texts or labels transfer.
- **Feature Transfer:** Linguistic knowledge is shared or transferred on the feature level between source and target languages.
- **Parameter Transfer:** Parameter values are transferred between language models. This effectively transfers the behaviour of the model from source to target languages.

Additionally, a comprehensive Table A.2 is provided to summarise all techniques utilised in our surveyed studies according to their publication year, cross-lingual transfer paradigm, models, description of the cross-lingual approach, and whether or not the codes and resources are available.

3.4.1 Instance Transfer

Instances in the cross-lingual hate speech task are comprised of two elements – texts and corresponding labels – in both source and target languages. Although the single source language data is not directly applicable in cross-lingual transfer, certain parts of the instance can still be repurposed together with target language data via instance transfer. Instance transfer (also called instance projection) revolves around the transfer of specific data elements (either text or label information) between source and target languages, which is a key technique in the realm of cross-lingual hate speech detection.

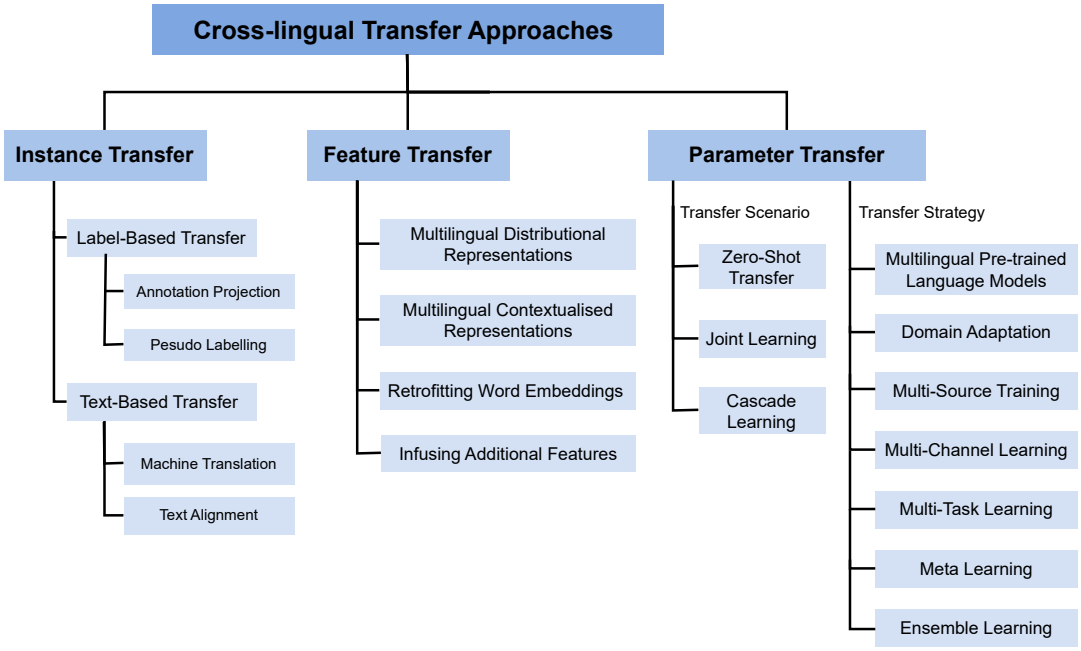

Figure 3.5: The hierarchy of cross-lingual transfer approaches.

To implement instance transfer, a correspondence between instances in the source and target languages must be created on the data level. *Correspondence* refers to the pair of instances with the same meaning of texts or identical labels [27], e.g., parallel data possess the same label matching, and translated text retains the same label of its original counterpart. Based on the established correspondence, texts or labels can be transitioned from one language to the other . Subsequently, these projected texts or labels can be used for further training and fine-tuning stages.

Pairs of instances in source and target languages that share identical labels are referred to as corresponding instances. All data-based transfer techniques necessitate the presence of corresponding texts/labels or a mechanism to generate them. Techniques for instance transfer are introduced distinctly for both text and label levels.

For label-based transfer, the primary strategies include annotation projection and pseudo-labelling. Meanwhile, For text-based transfer, the predominant approaches are machine translation and text alignment.

3.4.1.1 Label-Based Transfer

Annotation Projection

It involves leveraging parallel corpora to transfer annotations from a source language to a target language. Parallel corpora consist of source and target language data that are aligned at the sentence level, ensuring precise translations between each language pair, as opposed to machine-generated translations. Since data in the source language is paired with target language data, labels in the source language can be directly projected onto the target language. For example, if a sentence in the source language is labelled as hateful, its corresponding text in the target language can be inferred to have the same label. It can effectively create labelled data for the target language without manual annotation, especially useful when building datasets for languages where annotating from scratch might be challenging due to linguistic complexities or lack of expert annotators.

Pseudo Labelling

It is a semi-supervised learning technique, where a model trained on labelled source data is used to make predictions on unlabeled target data [135]. These predicted labels made with high confidence are treated as "pseudo-labels" for the unlabelled target data. By using high-confidence predictions from a model trained in the source language, it can generate labelled target data in another language. This augmented target data can then be used to train or fine-tune models for abusive language detection in the target language. To generate pseudo labels, some studies use an ensemble-based approach to generate labels based on majority voting for unlabelled target

datasets, and use bootstrapped target datasets to further fine-tune trained models by source training samples [123, 136, 137]. Hande et al. [124] directly use a pre-trained multilingual language model (i.e. XLM-R [36]) to predict the pseudo labels. Zia et al. [138] firstly fine-tune a pre-trained XLM-R on gold-labelled source language data, and then use it to create a new pseudo-labelled dataset in the target language.

3.4.1.2 Text-Based Transfer

Machine Translation

In the absence of sufficient annotated parallel data resources, machine translation provides us with an efficient alternative strategy to achieve text-level transmission. In general, it leverages translation tools [17, 120] or models [84] to translate labelled datasets between source and target languages, thereby promoting the augmentation of training data. Labels for translated data are the same as those in the original labelled dataset. Translation can be either directed or undirected: (i) target to source [43, 74, 139–141]; (ii) source to target [84, 142, 143]; (iii) translate both source and target to each other [29, 40, 144–146]. In addition, back-translation is commonly used as a data augmentation technique in cross-lingual scenarios by using various transformations on the original source data to create new samples for further model training [147]. Back-translation refers to translating a source sentence to a target language and then reverting it back to the original language, generating synthetic parallel data in the process. The dataset obtained through back-translation might have slight lexical variations but retain its offensive essence, thereby enhancing the model's robustness and generalisability.

Text Alignment

A process similar to machine translation, but instead of using translation tools or systems, existing mapping techniques between languages are employed to generate

labelled data for the low-resource language. Shi et al. [30] begin by using a shared space between two language vectors to identify the most similar sentence in the target language to the labelled data in the source language. The identified sentences are then treated as labelled data for the target language, and assigned the same labels that the source data possesses. Ryzhova et al. [148] augment training datasets by generating attacked source datasets. They apply an adversarial attack algorithm [149] to the source sentence by only replacing words that the model considers important, ensuring that the new sentence is semantically similar to the old one.

3.4.2 Feature Transfer

Feature transfer delves into approaches that operate knowledge transfer across languages at the feature level, transforming linguistic features to aid in the detection of hate speech content. Rather than directly translating source texts or projecting source labels, feature-level cross-lingual techniques focus on extracting salient linguistic features from source and target languages, and aligning them into a shared feature space, which ensures the essence of text (i.e. hate speech content) remains consistent across languages.

Cross-lingual Word Embeddings (CLWEs), also known as Multilingual Word Embeddings (MWEs), are commonly used on the feature level transfer in the cross-lingual scenario of hate speech detection. They train monolingual word embeddings (such as Word2Vec or FASTTEXT) on multiple languages. This yields vectors that can capture semantic similarities relying on shared representations between languages. For example, a word like "hate" in English and its equivalent "odio" in Spanish might be closer in this shared space, allowing for effective feature projection. As a result, a sentence can be represented with a similar set of vectors as its translations, thus a model trained on the source language may be applied to the target language without any intermediate transfer steps.

In a cross-lingual hate speech detection task, various approaches utilise existing CLWEs or retrofit CLWEs based on specific domains. Three main techniques for feature transfer are introduced: multilingual distributional representations, multilingual contextualised representations, and retrofitting word embeddings.

3.4.2.1 Multilingual Distributional Representations

Pre-trained distributional word embeddings are most commonly used as embedding layers in neural networks for different language data inputs in the cross-lingual hate speech task. Among these, MUSE embeddings have emerged as a predominant choice. They are extensively utilised to extract multilingual features and are often integrated with both traditional machine learning models, such as Gradient Boosted Decision Trees [74], and advanced deep learning architectures (i.e. LSTM [17, 40, 89, 120, 150], BiLSTM [29, 123, 136, 137], CNN-GRU [140, 141, 151], Capsule Networks [29], Sluice networks [82]), and BERT [17]. Other notable multilingual distributional vectors, such as Multilingual GloVe and FASTTEXT, have also been employed in architectures like BiLSTM [42] and Capsule Networks [29]. In addition, Babylon multilingual embeddings have been paired with Sluice networks for multilingual hate speech detection [82]. Bansal et al. [152] integrate bilingual switching features into the Hierarchical Attention Network (HAN) architecture, enhancing the transfer knowledge for cross-lingual detection.

Furthermore, some studies have ventured into cross-lingual projection using the MUSE mapping method, aligning monolingual FASTTEXT embeddings in English and German to produce their own CLWEs [123, 136, 137]. Paul et al. [153] align English and Hindi FASTTEXT embeddings by performing Canonical Correlation Analysis (CCA),[15] and project them into shared vector space where they are maximally correlated. Kapoor et al. [41] use multilingual datasets to train embeddings in order

[15] https://www.mathworks.com/help/stats/canoncorr.html

to capture specific distributional representations of tweets. De la Peña Sarracén and Rosso [154] propose a graph auto-encoder framework to learn embeddings of a set of texts in an unsupervised way, adding language-specific knowledge via Universal Sentences Encoder (USE).

3.4.2.2 Multilingual Contextualised Representations

Multilingual contextualised representations, typically derived at the sentence level from PLMs (such as mBERT [33] and XLM-R [36]), are able to capture deeper semantic and contextual relationships between words and phrases. Some researchers have leveraged hidden features with richer semantic information from the first embedding layer of mBERT [89, 120, 155], DistilmBERT [100, 124], XLM [156], XLM-R [124] as feature extraction layers. These embeddings are either extracted as standalone embedding layers [89, 100, 120, 155] or utilised as frozen embedding layers [124, 156] in deep neural networks, such as BiLSTM, yielding enhanced performance over traditional distributional embeddings.

Furthermore, multilingual sentence embeddings LASER has also found widespread applications across a variety of models, including Support Vector Machine (SVM) [155], Logistic Regression (LR) [120, 140, 141, 151], Random Forest (RF) [89, 155], XGBoost [89], Multi-Layer Perceptron (MLP) [157] and LSTM [100]. Rodríguez et al. [155] propose to use LabSE and mBERT representations with SVM-based and tree-based classifiers, achieving the best performance in cross-lingual hate speech research.

3.4.2.3 Retrofitting Word Embeddings

Retrofitting pre-trained word embeddings by infusing multilingual domain knowledge stands as a promising strategy to amplify the semantic relationships across languages, thereby enhancing the efficacy and scalability of models in cross-lingual abusive language detection. Pant and Dadu [158] employ a supervised FASTTEXT model trained

on the sarcasm detection corpus [159] to improve the identification of hate speech content in English and Hindi languages. Arango et al. [89] construct hate speech-specific word embeddings by aligning monolingual Word2Vec embeddings using hate-specific bilingual dictionaries (such as HurtLex), which can capture non-traditional translations of words between languages. Additionally, Hahn et al. [160] learn semantic subspace-based representations to model profane languages on both word and sentence levels. Jiang and Zubiaga [161] propose a cross-lingual domain-aware semantic specialisation system to construct sexism-aware word embeddings. They retrofit pre-trained FASTTEXT word vectors by integrating in-domain and out-of-domain linguistic knowledge (such as lexico-semantic relations) into the specialised feature space.

3.4.2.4 Infusing Additional Features

Additional features from various external resources are commonly used in identifying hate speech across languages, such as domain-specific [40, 160], language-specific [31, 89], and typographic features [162, 163]. Integrating these features can effectively enhance model performance for cross-lingual detection.

Specifically, domain-specific features are essential to assist the hate-related knowledge transfer. The multilingual lexicon, HurtLex, has been employed in multiple studies [17, 29, 40, 120]. Some researchers also utilise hateful word pairs to construct domain-aware or semantic-based representations across languages [89, 160, 161], due to the lack of hate-specific patterns in general-purpose multilingual embeddings. In addition, infusing language-specific features can be of great importance in integrating linguistic and cultural knowledge for geographically sensitive tasks, such as hate speech detection. It can be in different forms, such as language switching pattern matrix [152], bilingual pairs [161], cross-cultural similarities [89], and social dynamics among users [79]. This feature can provide a richer understanding of cultural dimensions in hateful content [31]. Furthermore, typographic features provide a language-

agnostic and domain-agnostic perspective to help identify online hate due to their shared meanings across languages, which includes capitals, punctuations, and emojis [134, 162, 163]. Stylometric features also remain persistent with respect to language variations, as the writing style of toxic content can be correlated to the emotional profile of social media users [163].

3.4.3 Parameter Transfer

The behaviour and performance of models in different NLP tasks are dependent on the values of model parameters. Parameter transfer assumes that some parameters or prior distributions of hyperparameters are transferred between different languages within one model or between individual models, and most parameter transfer approaches are inductive transfer learning in the cross-lingual hate speech detection task [28]. Inspired by Pikuliak et al. [27], we outline three distinct scenarios for all surveyed cross-lingual studies of hate speech phenomena, depending on how the source and target data are utilised during the process of parameter transfer, namely zero-shot transfer, joint learning, and cascade learning. Differences between these scenarios are highlighted in Figure 3.6. Only source data is used for training in zero-shot transfer, while both source and target data are applied during the training process in joint learning and cascade learning. Joint learning utilises both simultaneously to train the model, but cascade learning leverages them in the different stages of training.

Numerous parameter-level transfer strategies, tailored for the task of identifying cross-lingual hate speech, can be sophisticated and diverse in terms of their model architectures and training procedures. These strategies might encompass one or more of the model transfer scenarios introduced above. Table 3.2 has summarised the possible correlations between these transfer strategies and their application scenarios, and will discuss more details in the following subsections regarding scenarios and strategies utilised on parameter-level transfer.

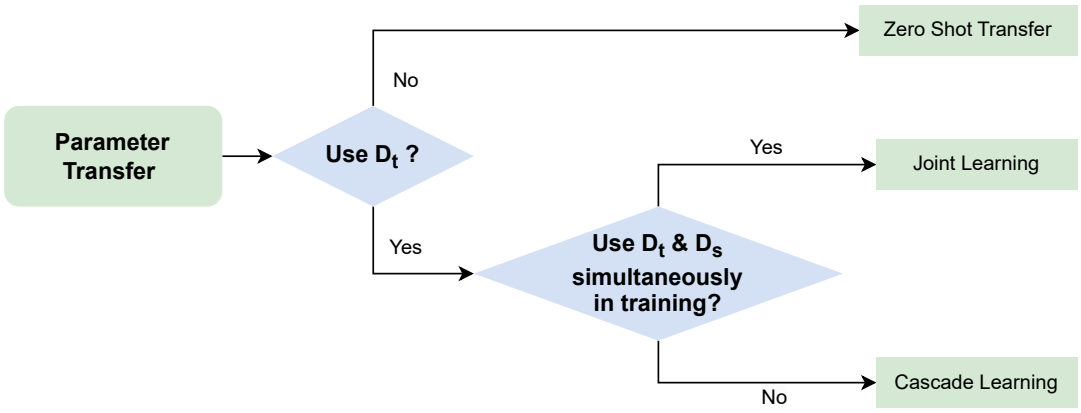

Figure 3.6: Different scenarios in parameter transfer for automated detection of cross-lingual hate speech.

Transfer Strategy	Zero-Shot Transfer	Joint Learning	Cascade Learning
Multilingual PLMs	✓✓	✓	✓✓
Domain Adaptation		✓	✓✓
Multi-source Training	✓✓	✓✓	✓
Multi-channel Learning	✓	✓✓	
Multi-task Learning	✓	✓✓	✓
Meta Learning		✓	✓
Ensemble Learning	✓	✓✓	✓

Table 3.2: Correlations between scenarios and strategies in parameter transfer for automated detection of cross-lingual hate speech. ✓✓means higher frequency than ✓.

3.4.3.1 Transfer Scenarios

We introduce three different scenarios that happened during the process of parameter transfer.

Zero-Shot Transfer

It is a promising scenario for cross-lingual hate speech detection, especially when labelled data in the target language is scarce. Zero-shot transfer refers to the scenario in which a model is able to detect hate speech in a language (target) it has never seen during training. Essentially, the model is trained on labelled source data only from one or more languages except for the target language, and is then expected to perform detection of hateful content on a completely different target language without

any labelled samples. Hence, no target data are used during the training process, and the whole model is transferred. This is sometimes called direct transfer or model transfer. There are 25 papers found in our survey that identify hate speech across languages in the zero-shot transfer scenario. Some subdivisions frequently emerge in the zero-shot transfer scenario.

- *Single Source to Single Target:* Models are trained on labelled data from a single source language and then directly adapted on an unseen target language [26, 31, 79, 84, 146, 155, 157, 164–172]. It is the most frequent approach (62.5%) observed in surveyed studies for zero-shot transfer and is considered a subset of the domain adaptation strategy (described in Section 3.4.3.2).

- *Multiple Sources to Single Target:* Models are trained on labelled data from multiple source languages (exclude the target language) and then evaluated on a single target language [140, 148, 166]. It belongs to a subset of multi-source training strategy that the combination of training datasets excludes the target data (described in Section 3.4.3.2).

- *Pseudo-Target Augmented Training:* The source data is translated into the parallel target data via machine translation tools, and then models are trained on both the original source data and the translated target data together [156] or in parallel [17, 29, 40, 120]. It falls under the multi-channel learning strategy when using both source and translated target data in parallel[16] and fusing them before the output layer of the model. This allows the model to learn nuances specific to the target language from the pseudo content.

- *Parameter Frozen:* A model is trained on source data first, and then a subset of its parameters (e.g., weights from specific layers) is saved to initialise an-

[16]Pseudo-Target Augmented Training has overlapped with the joint learning scenario. It distinguishes itself by using pseudo target texts for training instead of real target texts.

other model for target language evaluation [83, 171, 173, 174]. It is a subset of the domain adaptation strategy,[17] and captures language knowledge from certain layers of the source model to provide a head start for the target model, potentially leading to faster convergence and better generalisation.

It is worth highlighting that these methods show the reliance on multilingual PLMs, especially often used for Single Source to Single Target and Multiple Sources to Single Target scenarios. mBERT and XLM-R are two of the most prominent multilingual PLMs in surveyed papers, because of their popularity, widespread adoption and SOTA performance in most multilingual NLP tasks. They together account for 80.8% of papers in the zero-shot transfer scenario. Additionally, some papers release domain- and language-specific pre-trained model for multilingual [168], choose other multilingual PLMs like XLM [156] and language-specific MuRIL [146, 168], or combine multilingual word embeddings with monolingual neural models like LSTM [29, 40] and BERT [17].

Joint Learning

It typically involves training a model between languages based on multiple channels, tasks or objectives simultaneously, or jointly assimilating multiple types of data or modalities. The model might have some parameters that are shared across languages as well as some that are specific to each, so it is also called *parameter sharing* [27]. The transfer of knowledge happens only on shared parameters. Changes to the shared parameters can affect each other between languages. They are interrelated and can provide complementary information.

In this scenario, both source and target data are used at the same time during the training stage. Parameters can be shared between source and target channels

[17]Parameter frozen in zero-shot transfer scenario shares similarities with the cascade learning scenario. However, while cascade learning continues to fine-tune the trained model with target data, zero-shot transfer directly applies the trained model to test the target data without further fine-tuning.

within one model or between individual source and target models. We discovered 23 papers that leverage joint learning strategies for cross-lingual hate speech knowledge transfer. The distinction among these strategies mainly lies in the stage of knowledge transfer across languages and the extent of parameter sharing (either all or subsets of parameters to share) [162]. Based on these criteria, we categorise them into three primary fusion stages: early fusion, model-level fusion, and late fusion (as shown in Figure 3.7).

- *Early fusion:* A combination of datasets or extracted embeddings from both source and target languages, with the entire model sharing parameters. It is commonly adopted in few-shot learning [31, 146, 156, 164, 168, 171] and multi-source training [42, 143, 146, 151, 163, 168, 169, 172, 175, 176], where datasets from the target language are incorporated during training. In few-shot learning, the model is trained by utilising a very small amount of target samples combined with a more extensive source dataset, while multi-source training typically takes one mixture dataset of source and target languages or integrates multiple datasets including multiple non-target datasets with the target dataset for training.

- *Model-level fusion:* A concatenation of high-level hidden features within a model, and only specific layers or components of the model share parameters, facilitating a more nuanced transfer of knowledge. It commonly occurs in parallel model architectures, such as multi-channel learning [144, 162], multi-task learning [82, 177] and meta-learning [178, 179].

- *Late fusion:* A combination of different channels or individual models before the output layer, and only the parameters of the output layer for predictions are shared, ensuring a final consolidation of knowledge from various languages. It is often applied in multi-channel architectures [43].

Early fusion is the most prevalent fusion strategy due to its straightforward integration of source and target language data for better generalisation, while late fusion provides a modular approach, allowing for the independent training of individual models that can later be merged. However, unlike early fusion, model-level fusion effectively overcomes the curse of high dimensionality and the synchronisation demands among diverse features. Moreover, it ensures that interactions between different languages are not isolated as in late fusion [162], showing its robustness and resilience.

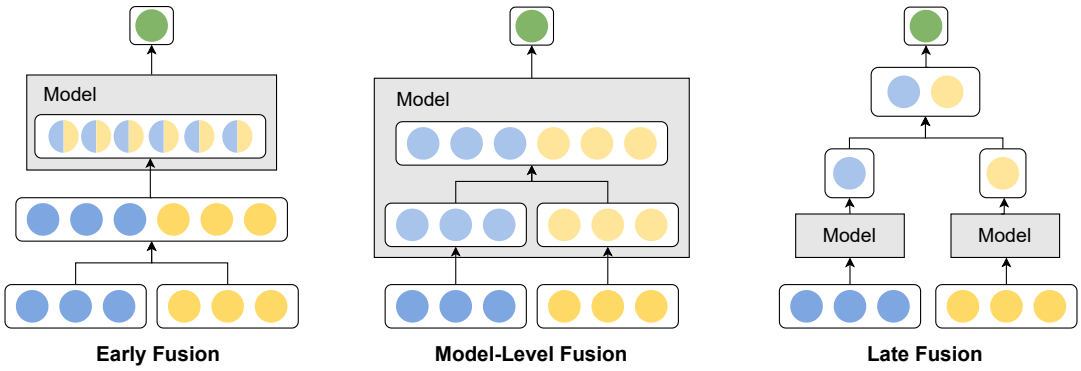

Figure 3.7: Different fusion stages in the joint learning scenario.

Cascade Learning

In the cascade learning scenario, both source and target data are used at different times during the training. First, an existing pre-trained model is directly used or a new model is trained with source data, and then the trained model is fine-tuned with target data, where subsets of parameters are shared between different stages of training and fine-tuning. Similarly to joint learning, various strategies can be applied to training and fine-tuning stages according to 19 surveyed papers.

A significant portion of research directly employs existing SOTA pre-trained multilingual models serving as the foundation,[18] while some studies also initiate to train

[18]Cascade learning fine-tunes PLMs using the target data, while zero-shot transfer fine-tunes

a model with one or multiple source datasets. Subsequent to this initial training with source data, the fine-tuning stage in cascade learning is applied across an entire [41, 141, 164, 180, 181] or only a limited subset of target data (i.e. few-shot learning) [169, 180], optimising the model adaptability. Röttger et al. [99] indicate that an initial training phase on source data (English) could increase model performance when there is little fine-tuning data in the target language, where source data can partly substitute target data in few-shot settings. Some researchers delve into consistent fine-tuning [148, 178, 179] or iterative fine-tuning (i.e. meta learning) [178, 179], performing multiple rounds of fine-tuning on the pre-trained model to further enhance its ability to transfer knowledge.

Additionally, given the limited availability of target data, data augmentation techniques have emerged to address this challenge. They aim to enrich the target dataset to enable fine-tuning stage, by enhancing label diversity or expanding the text corpus. Some studies have attempted to generate pseudo labels for unlabelled target datasets [123, 136–138] (as Section 3.4.1.1 mentioned), while others augment and refine pseudo texts by exploring advanced techniques, such as word alignment to project samples [30], adversarial algorithm to generate attack samples [148], and domain-specific target data filtering [84].

3.4.3.2 Hybrid Transfer Strategies

This subsection delves into a range of diverse transfer strategies, each extensively employed to tackle the complicated challenges of cross-lingual hate speech detection. According to the utilisation of source and target data, these strategies can be broadly covered in one or more parameter transfer scenarios mentioned in Section 3.4.3.1. That is, these strategies are not mutually exclusive. A single scenario can seamlessly integrate multiple transfer strategies (see Table 3.2). This collaborative transfer ap-

PLMs using the non-target data.

proach ensures comprehensive and effective knowledge transfer throughout the whole process.

Multilingual Pre-Trained Language Models

These models are regarded as a cornerstone of parameter-level transfer techniques for low-resource scenarios, especially in the field of cross-lingual hate speech detection. They are trained on large-scale multilingual corpora, learning commonalities and differences between languages. Therefore, a single pre-trained model can share the same underlying sub-word vocabulary and semantic representations across languages [33]. Multilingual PLMs can be initialised as the base model with pre-trained parameters, then fine-tuned on available hate speech data in all three scenarios of parameter transfer (zero-shot transfer, joint learning and cascade learning) [164, 172]. Beyond this, they can also employed in the other two transfer levels (data and feature). They are able to augment target datasets by predicting pseudo labels for unlabelled data and generating synthetic target texts. And they can be used as multilingual representations, where subsequent models are constructed based on these source representations [155].

Domain Adaptation

It involves taking a model trained (or pre-trained) in the source language and adapting its parameters to initialise a new model for the low-resource language, aiming to leverage source data to improve performance on target data [84, 182]. Its advantage is to capture hate speech knowledge and linguistic similarities between the languages within the model's parameters. In the zero-shot transfer scenario, parameters of the trained model are entirely or partially used on the target data for prediction [171, 173]. In addition, cascade learning adopts a more subtle approach. That is, the entire or specific subsets of the model's parameters (like specific layers) are frozen to retain

their original state [169], while others are fine-tuned to better align with the target data, often using a smaller target dataset (as few-shot learning) [99, 180].

Multi-Source Training

This approach often entails the amalgamation of labelled data from multiple source languages and potentially the target language in diverse ways. By synthesising information across multiple languages, this strategy can capture the diversities and commonalities of linguistic patterns, demonstrating significant efficacy in the cross-lingual setting. It is extensively utilised across all three transfer scenarios, serving as an augmented cross-lingual learning approach during different stages of training or fine-tuning. In multi-source training, the most prevalent way is to train or fine-tune the model on multiple source languages. This can be executed in the training phase of joint learning [42, 143, 145, 146, 151, 163, 172, 175–177, 183] or the fine-tuning phase of cascade learning [168, 169] in conjunction with the target language, or in isolation from the target language for zero-shot transfer [148, 166]. Deshpande et al. [151] offer a more focused perspective by narrowing the linguistic scope to specific language families. By training models on languages within one language family, they then assess the model's performance on each member of that family, revealing the interplay of linguistic knowledge within related languages. It is worth noting that merging training examples from different languages can be detrimental to cross-lingual performance in cases where hate speech domains are too distant [84].

Multi-Channel Learning

This strategy typically refers to the use of multiple types of input representations or sources concurrently for a single task, where multiple embeddings or pre-processed versions of the text are often channelled as parallel input streams. In this architecture, each channel processes the input independently through parallel layers. These

independent processing channels are eventually combined based on their outputs for subsequent learning or classification stages. Such a design ensures that diverse input representations are holistically integrated, capturing a broader spectrum of linguistic knowledge across languages. Multi-channel learning finds its applications mainly in the zero-shot transfer and joint learning scenarios. When applied to cross-lingual hate speech detection, this approach often treats datasets from different languages as distinct input channels. Each channel adopts different neural networks or PLMs based on diverse input languages, and these channels are then merged into the final classification stage [144, 156, 162]. A significant improvement to this strategy is the incorporation of machine translation techniques. By translating labelled source data to pseudo target data, low-resource target data is enriched and a bridge is constructed between source and target languages, enhancing knowledge relationships and a more seamless transfer of languages [17, 29, 40, 43, 120].

Multi-Task Learning

It involves training a model on several interrelated tasks simultaneously, where the proficiency learned from one task can boost performance on another, thereby creating a joint learning environment. A key point of this strategy is the shared representation of hidden features across specific layers of the model. Such shared layers enable different tasks to mutually benefit from common representations, while also preserving their distinctiveness through task-specific output layers. The training process is holistic, aiming to optimise the model's performance across all tasks. This is typically achieved by incorporating the loss functions from each task. Multi-task learning emerges as a potent technique for cross-lingual hate speech detection within the parameter-level transfer techniques across three transfer scenarios. Models can be trained on the hate speech detection task and a wide range of related auxiliary tasks, such as Sentiment Analysis, Named Entity Recognition (NER), Dependency

Parsing, and Part-Of-Speech (POS) Tagging [82, 166, 177]. Additionally, they can be fine-tuned on data in various offensive categories, such as aggressive content [83, 173]. This multi-task training strategy enables the model to identify and process offensive content with higher accuracy and enhances the model's ability to grasp linguistic nuances across languages, thereby improving the performance of cross-lingual hate speech detection.

Meta Learning

This strategy, often referred to as "learning to learn", has become a popular few-shot learning technique in NLP. It involves training models to learn the optimal initialisation of parameters, allowing them to be fine-tuned with a small amount of data from the target language. This technique finds its application within the joint learning and cascade learning transfer scenarios. It enables rapid adaptation to new unseen languages, making it especially useful in scenarios where labelled data is limited. Hence, this technique is critical for cross-lingual hate speech detection, especially for low-resource languages. The main meta learning methods include optimisation-based [184], and metric-based techniques [185]. Two papers found in the review employ optimisation-based meta learning frameworks. Montariol et al. [166] first to study meta learning for the problem of few-shot hate speech detection in low-resource languages. They propose a cross-lingual meta learning-based approach based on optimisation-based Model-Agnostic Meta-Learning (MAML) and Proto-MAML models, fine-tuning the base learner XLM-R with parallel few-shot datasets in different target languages. Awal et al. [179] propose HateMAML, a model-agnostic meta-learning-based framework that uses a semi-supervised self-refinement strategy to fine-tune a better pre-trained model for unseen data in the target language, showing effective performance for hate speech detection in low-resource languages.

Ensemble Learning

It leverages multiple models or "learners" to collaboratively make decisions. Instead of relying on the output of a single model, ensemble learning combines all predictions from multiple machine learning or deep learning models to produce a final prediction based on majority voting [153, 163]. It takes advantage of the different strengths of each model, mitigating the weaknesses and biases of any single model and avoiding overfitting, improving robustness and reliability over a single classifier. This is particularly beneficial in cross-lingual settings, where language biases can pose challenges to individual models. In cross-lingual hate speech detection, ensemble learning is mainly applied in the zero-shot, joint learning, and cascade learning transfer scenarios. Deep ensemble models have emerged as the predominant choice in recent studies [123, 136, 137, 153], while the integration of PLMs within ensemble learning frameworks has also presented a significant performance [148, 150, 163]. Each model in the ensemble can be trained on different languages or linguistic features, allowing the ensemble model to capture diverse linguistic patterns and reduce test errors for target languages [148].

3.4.4 Summary of Cross-lingual Approaches

For transfer scenarios, zero-shot transfer does not use any target data, while joint and cascade learning do. With zero-shot learning, we always transfer the whole model, while with joint and cascade learning we might transfer only a subset of parameters. Some overlaps of transfer strategies might exist between each transfer scenario because multiple strategies can be conducted based on more than one transfer scenario.

Model distributions. We provide an overview of diverse models and their frequency used in recent cross-lingual studies (see Figure 3.8), which reveals a clear trend towards more advanced architectures. Traditional machine learning-based models have been employed 24 times, while deep learning-based models have seen a slightly

higher usage (38 times) due to their complex structure and learning capacity from a large amount of data. Additionally, given that PLMs encompass multilingual knowledge, transformer-based models have been overwhelmingly preferred in relevant cross-lingual studies (99 times), highlighting their capacity to address complicated linguistic challenges and their growing dominance in the current NLP research. Large Language Models have also recently been investigated in this area due to their excellent emergent ability.

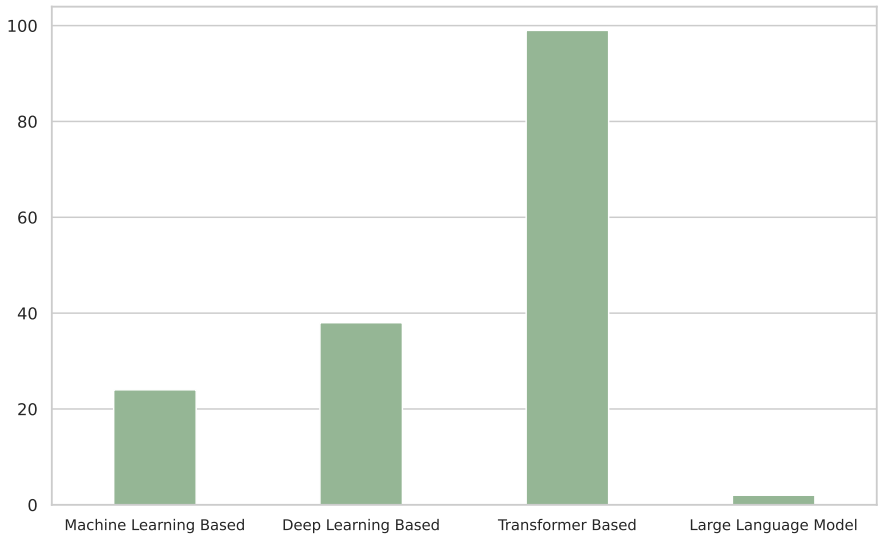

Figure 3.8: Number of different types of models used in the surveyed papers for cross-lingual hate speech detection.

Evaluation metrics. To evaluate model performance in cross-lingual hate speech, macro-averaged F1 score is the most used metric to due to class imbalances in existing hate speech datasets, while accuracy is also popular. Some studies also consider precision, recall, and weighted F1 score for a detailed assessment. Eronen et al. [167] propose a new linguistic similarity metric based on the World Atlas of Language Structures (WALS) [186] to select optimal transfer languages for automatic hate speech

detection.

Source code availability. Out of a total of 67 papers reviewed, 26 papers have made their source code available online (approximately 38% of papers), indicating a positive trend towards open-source sharing. However, a significant portion, 41 papers (61.2%), do not yet provide access to their code. Making source code available could provide more practical solutions to the task of cross-lingual hate speech detection, facilitating transparency and reproducibility of research.

Language pairs. We present the frequency of language pairs in cross-lingual studies in Figure 3.9. English shows the predominant use as the source language, while the most frequent target languages are Spanish, German, and Italian. These languages share typological similarities with English, which likely facilitates their frequent pairing in cross-lingual hate speech detection. This emphasises the impact of data availability and language similarity on the training and application of cross-lingual methods, while also revealing the limitation of relevant cross-lingual research resulting from data scarcity. Many studies have also explored cross-lingual learning by transferring English to Arabic, Hindi and French, as these languages have a relatively large amount of datasets available (as presented in Figure 3.2). Additionally, some studies working on the transfer between non-English languages, among which Danish, Indonesian, Portuguese, and Turkish show higher frequencies.

3.5 Current Challenges

In this section, we examine the primary challenges encountered in cross-lingual hate speech detection till 2023 within NLP field into three aspects (language, dataset and approach), highlighting the complexity and limitation of cross-lingual tasks.

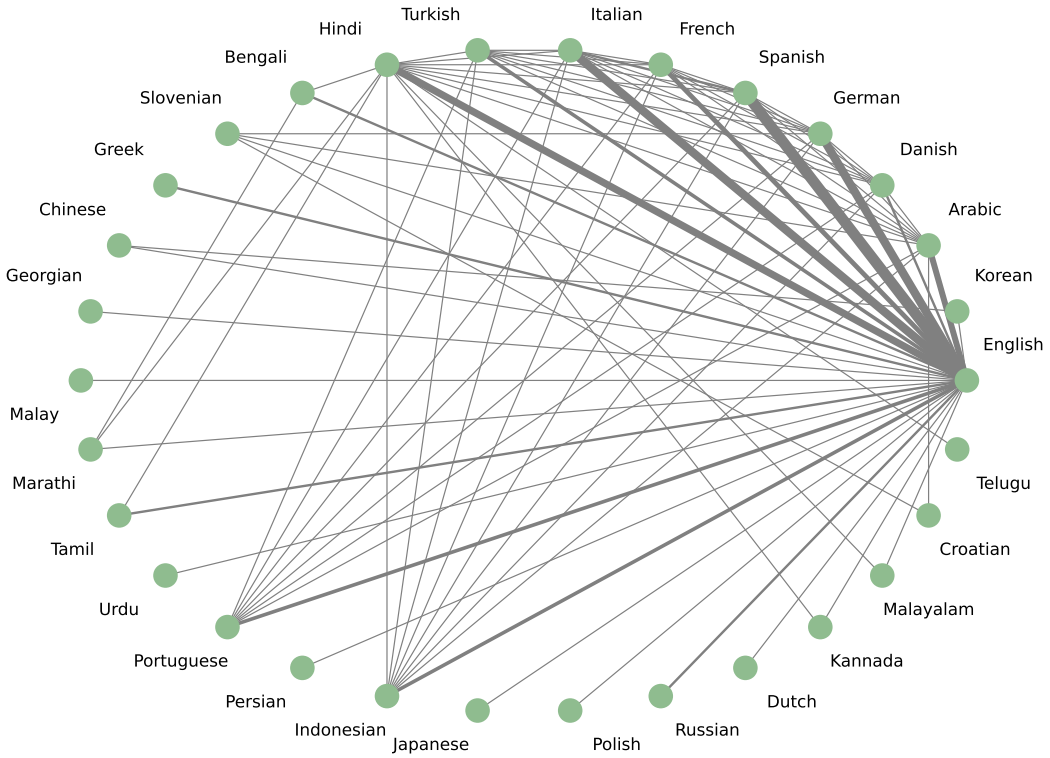

Figure 3.9: Synergies between languages. A link between two languages indicates that both have been used simultaneously in a model, as a source or target language, or to learn multilingual feature spaces. Higher frequencies correspond to thicker lines.

3.5.1 Language-Related Challenges

Diverse Linguistic Structures

The inherent diversity of linguistic structures across languages and their relationships within language families or dialects poses a significant challenge to detecting cross-lingual hate speech. This diversity encompasses various aspects from basic grammatical rules to complex nuances unique to each language. Since languages within a certain language family may exhibit similarities, cross-lingual transfer works better in more closely related languages and poses challenges for more dissimilar languages [183]. However, this does not necessarily imply a consistency in linguistic features

[167]. Furthermore, the presence of dialectal forms within a single language amplifies its complexity, complicating the detection task. For example, the Arabic language is spoken by a wide range of countries across Asia, which creates complex forms by its various dialects [181].

Code Mixing

Non-English social media content is often characterised by code-mixing, where users blend languages, use transliterations, or incorporate multiple scripts within a single sentence or conversation, such as English-Hindi [43, 124] and English-Chinese [161]. The English-Hindi combinations (also called Hinglish) appear in many cases, because the unique policies on regulating such speech in these regions add layers of complexity and make detection more challenging [41].

Cultural Variations

Culture is multifaceted and complex. The perception of hate speech can differ significantly across languages and cultures [79]. What is considered non-offensive in one culture or language might be misinterpreted as signals of offence when transferred to another, leading to detection discrepancies. Even within a single language, there can be vast cultural diversity. For example, categorising English as a representative of "western cultural background" might ignore the differences between American and British cultures [31]. Such cultural variances can hinder the effective transferability of language models [183]. Moreover, hateful expressions often employ figurative language, rhetorical figures and idioms, which are also language-dependent. This requires models to interpret underlying intent rather than relying solely on literal meanings [120].

3.5.2 Dataset-Related Challenges

Limited Labelled Datasets

One of the main challenges in cross-lingual transfer is the lack of adequately labelled datasets, especially for low-resource languages. These languages often do not have the extensive presence or attention that major languages receive, leading to limited data availability for training and evaluation [150] and limited studies in detecting non-English hate speech (as shown in Figures 3.2, 3.3 and 3.9). Additionally, creating a high-quality labelled dataset requires extensive manual effort, which can potentially lead to high annotation and time costs as well as be harmful to annotators [99]. Annotating hate speech content is even more challenging due to the need for cultural and linguistic knowledge. Ethical and privacy concerns also arise when dealing with real-world data from social media platforms or other online sources [20], because there are potential risks of exposing sensitive information or inadvertently promoting hate speech when studying it or trying to combat it.

Inconsistent Definition of Hate Speech

Hate speech encompasses a broad spectrum, from misogyny and racism to other forms of discrimination, often leading to overlaps between datasets [79]. Labelled corpora vary in the general or specific subtypes of hate speech. This could make the available resources in low-resource languages even scarcer [123, 136, 137], and introduce ambiguities that challenge the hate speech detection task, especially in multilingual scenarios [100]. The lack of definition consistency across datasets strongly limits cross-lingual research on hate speech, because many datasets in either source or target languages may be incompatible for use in combination [137].

Inconsistent Annotations

Hate speech datasets can easily exhibit systematic gaps and biases due to how they are annotated [183]. Different labelled datasets often utilise distinct annotation frameworks or strategies [1]. The subjective nature of hate speech leads to tentatively disputed annotations among human annotators, where hateful content could be interpreted differently depending on annotators' understanding [146]. Annotating these contents is easily influenced by individual demographics and cultural backgrounds, often resulting in low inter-annotator agreement [163]. Additionally, there is an "annotation's dilemma" [141], whereby ambiguous instances might be annotated incorrectly by annotators but predicted accurately by the model. The inconsistent annotation and such errors further complicate the task of training robust hate speech detection models, especially when applying CLTL approaches.

Dataset Imbalance

The datasets often exhibit imbalances that can adversely affect cross-lingual transfer performance. One primary concern in a cross-lingual setting is class imbalance. Given that the majority of social media content is non-hateful, the label distributions in datasets are typically skewed towards the non-hate label [123, 137, 142]. This skewness can lead to training issues, especially when working with small training corpora. In addition, when considering multilingual or merged datasets, inter-language imbalance is another challenge [151]. Such datasets often display significant distinctions in the number of examples available for each language. This imbalance can unintentionally bias models towards the semantic tendencies of languages that are overrepresented.

Dataset Bias

In addition to variations in definitions, annotations, and class distributions, biases in datasets can be multifaceted, such as topic, authorship, political affiliation, social

media platform, and collection period. These biases can be one of the major issues and lead to potential generalisation challenges for cross-lingual models. Among them, topic bias has been taken seriously in many works [74, 99, 100, 183]. While some datasets are general hate speech, others might concentrate on specific topics such as immigrants, misogyny, politics, or religion. Such topic-specific biases across datasets can harm model performance in cross-lingual classification [120, 156]. Besides, the temporal aspect of data collection, affected by events in different periods, can further introduce bias, leading to datasets with varied topical focuses [156].

Data Source Obsolescence

The rapidly evolving user-generated content on social media presents a significant challenge in cross-lingual hate speech detection. As online language rapidly transforms, especially influenced by moderation and real-world events, datasets can quickly become obsolete [100]. This obsolescence is further exacerbated by the emergence of new slang, metaphors, and colloquialisms that vary across languages and regions. For instance, terms that normally have neutral meanings (such as donkey) may be used offensively in sexist text [161]. The constrained and informal style of tweets, which are more like oral expressions than written language, complicates the preprocessing steps and leads to erroneous predictions [17, 141]. Emojis, associated with various forms of online harassment, further pose unique challenges [134]. Moreover, the domain and target of hate speech can shift significantly over time. Real-world events, from local incidents to global crises like the COVID-19 pandemic, can give rise to new domains and terms of hate speech [166]. Then, it becomes increasingly challenging to create and continually update datasets customised for every possible language and domain, leading to frequent low-resource issues.

3.5.3 Approach-Related Challenges

Limited Ability of Multilingual Pre-Trained Language Models

Cross-lingual models, especially multilingual PLMs, often face challenges related to their generalisation abilities [160, 180]. The generalisation ability of transformer-based multilingual PLMs can be inconsistent, especially for typologically diverse languages, because they might be pre-trained with a highly focused set of languages or some languages with insufficient training examples [151]. This can lead to bias and instability in model performance, depending on the model architectures or topical focus in datasets [120, 178]. Model overfitting across target languages is another important issue. Although multilingual models are less prone to overfitting on dataset-specific features than monolingual models, they will achieve poorer performance than monolingual ones in higher-resource settings, and they may require very different calibrations and adaptations across languages [99, 183].

Cross-lingual Transfer Performance

CLTL, the process of applying knowledge learned from one language to another, has emerged as a promising solution and has indeed shown effectiveness in hate speech detection, particularly when addressing the data scarcity issue [137, 144]. However, hate speech is deeply rooted in the specificity and diversity of language and culture, and its complexity poses significant obstacles to cross-lingual transfer, especially in zero-shot settings [79, 166]. Zero-shot transfer approaches to multilingual training often suffer from performance deficiencies when compared to models trained on actual target language data [180].

Limited Machine Translation

Machine translation plays an important role in cross-lingual studies as a strategy to augment datasets and alleviate data scarcity issues. A prevalent practice is to utilise Google Translate API.[19] [29, 120, 141, 145] However, the effectiveness of cross-lingual detection models is inherently related to the quality and accuracy of the machine translation tools. True semantics of the target text may change during the translation process. Machine translation tools may inadvertently diminish the toxicity degree and thereby reduce the perception of hateful or offensive content, especially in context-sensitive situations [146]. Therefore, models that rely on such translations can experience performance degradation, especially when predicting instances with semantic shifts. However, despite translation errors and uncertainties, they can still be valuable supplementary inputs for text classification tasks [144].

Poor Model Interpretability

Given the sensitive nature of the hate speech detection task, the interpretability of models is of prime importance. While advanced deep learning-based models have shown good performance in detecting hate speech, they often operate as "black boxes" with a lack of transparency in the decision-making process [141]. This opacity makes it hard for humans to understand the underlying reasons for the model performance in a particular language or scenario, and analyse the model errors [100, 151]. To bridge this interpretability gap, some tools are utilised to explain what models are doing, such as Local Interpretable Model-agnostic Explanations (LIME) [187] and SHapley Additive exPlanations (SHAP) [188], and theoretical approaches from cognitive linguistics are also applied like of frame semantics [100]. Human-in-the-loop paradigm can also help by integrating human expertise into the model's learning process, but it can be difficult and uncertain for individuals to provide effective feedback to enhance the

[19] https://translate.google.com

model performance [100].

3.6 Conclusion

In this chapter, we present the first systematic and holistic overview of recent studies using CLTL to detect offensive language in social media across languages. It contains 67 papers, describing them according to diverse aspects, including the multilingual datasets employed, cross-lingual resources leveraged, levels of transfer, and cross-lingual strategies applied. In addition, we present current challenges as well as two comprehensive tables in Appendix A containing multilingual datasets and CLTL techniques used in surveyed papers respectively to facilitate easy comparison and discovery of related works.

4

Collection of Sexism Dataset and Lexicon in Chinese

In this chapter, we address the problem of the scarcity of Chinese resources in the field of hate speech especially for gender-related content in social media. We define a methodology for the collection and annotation of Chinese online sexism at different levels of granularity, providing the first such effort in Chinese in sexism and hate speech. A Chinese sexist lexicon is created to assist research in Chinese sexism detection. Both the dataset and lexicon are then utilised to evaluate the effectiveness of existing state-of-the-art (SOTA) models in detecting Chinese sexist content.

The chapter is organised as follows. In Section 4.2 we describe the process of collecting and organising source data from Sina Weibo. Section 4.3 presents guidelines and evaluation of three annotation tasks for the collected dataset. The procedure of building a sexist lexicon is introduced in Section 4.4. Then we describe experimental results and analysis for sexism detection in Section 4.6. Section 4.7 discusses potential areas of research enabled by our dataset and lexical resources.

4.1 Introduction

When it comes to sexism-related datasets and resources, however, most efforts have been made for Indo-European languages [14, 46, 61, 63, 64], while the development of Chinese sexism identification is hindered due to the lack of Chinese annotated resources and Chinese sexism-related lexicons. Moreover, the creation of such resources poses several challenges when it comes to data collection and annotation, especially with the diversity of Chinese dialects and the ambiguity brought about by the emerging Internet language.

Hence we investigate how diverse behaviours, beliefs and attitudes towards women are expressed in social media, and focus on collecting data resources about sexism in Chinese. Given the modest presence of Chinese content and geographical access restrictions on Twitter, here we focus on the most prevalent microblogging platform in China, Sina Weibo. As a platform integrating the major features of Twitter, Facebook,

and Instagram, users of Sina Weibo can share posts (weibos) with texts, photos, and videos, which can trigger replies between users (comments) and endorsement (likes) from others [189, 190].

By using Sina Weibo to collect sexism-related weibos and comments, we build, annotate and analyse the Sina Weibo Sexism Review (SWSR) dataset. The SWSR dataset consists of two parts: *SexWeibo* and *SexComment*, both of which include the textual content of posts along with anonymised information of users, number of likes and other metadata. The process led to a dataset with 1,527 weibos and 8,969 comments. In addition, to assist research in the detection and analysis of sexist comments in Chinese, we provide a sexism-related offensive lexicon SexHateLex which aggregates and extends existing lexical resources in Chinese. Furthermore, we present the first experimentation in Chinese sexism detection to provide a benchmark, including the implementation of various machine learning and deep learning methods. Our experiments and methodology for sexism detection aim to further research this task in Chinese, as well as enable similar research efforts in other types of hate speech detection in Chinese. Our Chinese dataset and lexicon also enable multilingual sexism research which breaks the restriction of limited language resources. Abundant demographic and Weibo-based features in SWSR empower to exploit relevant studies on online abusive language in different aspects.

4.2 Data Collection

In describing our data collection process, we first describe the key characteristics of the Sina Weibo microblogging platform we use to build our SWSR dataset, discussing the different data harvesting options across the different weibo platforms. Then we delve into the data collection and filtering process.

4.2.1 Sina Weibo

Sina Weibo is the largest microblogging service in China, which has some unique characteristics with respect to Twitter. It is aimed at information sharing, dissemination and information acquisition based on user relationships [189]. Content on Sina Weibo is spread through the "following-follower" networks established between people [191], for example, allowing users to post comments on someone's Weibo or to reply to other people's comments on someone's Weibo. It allows users to insert images, videos, music, long articles and polls.

Sina Weibo has three main ways of accessing its website, namely weibo.com, weibo.cn and m.weibo.com. We can access Sina Weibo via PC terminal through weibo.com and weibo.cn, and the mobile counterpart is m.weibo.com. The weibo.com is more complex than weibo.cn because its Weibo page presents a richer functionality with more components which weibo.cn does not have, such as Top Topic Ranking, Hot Movie Recommendation, advertisements, etc. However, we can see in an example of weibo.cn in Figure 4.1 that the website structure is simple and straightforward. Both the weibo and its associated comment list can be easily retrieved and parsed for data collection. So we finally decide to use **weibo.cn** as the source website of Sina Weibo.

4.2.2 Data Collection and Processing

As described above, a Sina Weibo timeline comprises posts (weibos) which receive replies (comments). Initially, we use a keyword-driven method to collect a set of weibos, for which we then collect the associated comments. While the collection of weibos is restricted to those containing the keywords, our focus on the associated comments allows us more flexibility, retrieving content which need not contain the seed keywords. Figure 4.2 shows an overview of the data collection process, which we introduce further details in the steps below. Our SWSR dataset therefore is made of two tables for weibo and comment data along with some anonymised user information

Figure 4.1: An example of Sina Weibo on weibo.cn

pertaining to the weibos and comments. This user information includes features such as user gender and user location. All personally identifiable information is removed and not disclosed, including user names and mentions.

Step I: Extract Weibo Data

To construct our dataset, we use keyword-driven search to collect gender-related weibos from Sina Weibo platform (weibo.cn). In terms of relevance to the topic and through manual exploration [23, 61], we firstly determine to use seven different keywords for weibo data collection, namely 婊子 (bitch), 女同性恋 (lesbian), 女权 (feminism), 厌女 (misogyny), metoo 运动 (metoo movement), 性别歧视 (gender discrimination) and 性骚扰 (sexual harassment). We search and extract weibos containing these keywords. These keywords revolve around the core theme of sexism, covering a variety of levels from individual insults to widespread social movements, reflecting the current society's attention to gender issues. They involve not only negative

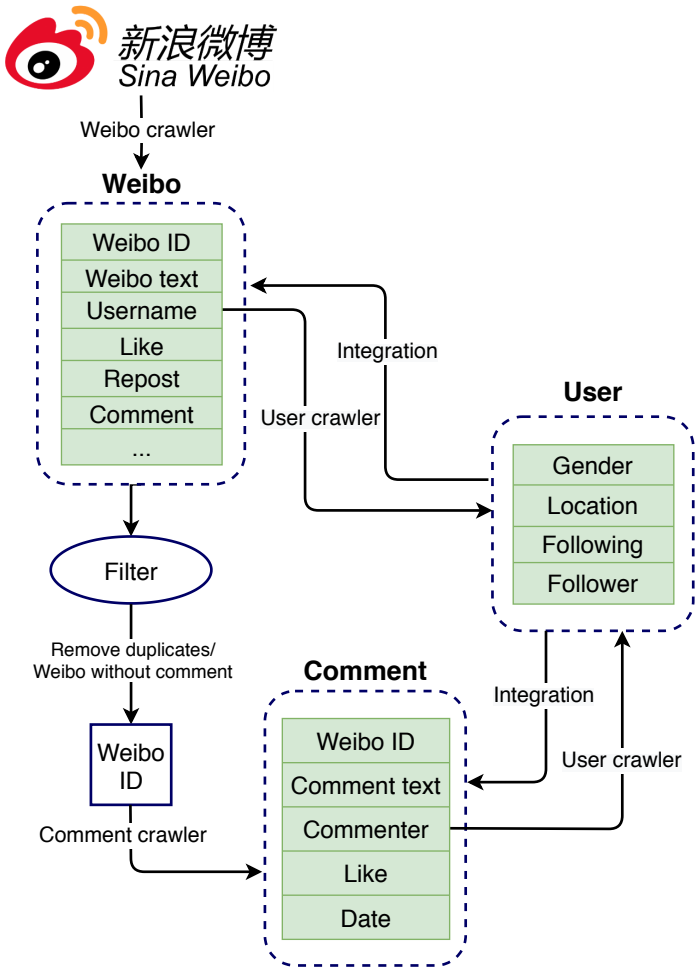

Figure 4.2: Overview of the data collection process.

stereotypes and prejudices against women, but also the challenges faced by the lesbian community, the progress of the feminist movement, the expression and impact of misogyny, and the "MeToo movement" that has sparked widespread discussion in recent years. In addition, they also involve the specific manifestations and far-reaching impacts of sexism and sexual harassment in daily life. Through these keywords, we strive to cover issues related to sexism for collected weibos, ensuring their diversity and complexity.

In addition, we retrieve user profiles, which include self-reported values such as gender and location, and other variables such as the number of followers. To protect user privacy in the dataset, usernames are anonymised by replacing them with a special token <username>. Then we combine these features into the weibo. The number of weibos collected for each keyword is listed in Table 4.1, which amounted to a total of 9,087 weibos collected for all keywords. Data collection was limited to posts made between June 2015 to June 2020.

Table 4.1: Number of weibos collected for each keyword.

Keyword	Translation	Number of Weibos	Total
婊子	bitch	407	
女同性恋	lesbian	520	
女权	feminism	2255	
厌女	misogyny	1757	**9087**
metoo 运动	metoo movement	1340	
性别歧视	gender discrimination	1366	
性骚扰	sexual harassment	1442	

Step II: Process Weibo Data

In this step, we process the collected weibos before collecting the associated comments in subsequent steps. We remove the weibos that match at least one of the following criteria:

- Weibos without any comments. This can be easily done by checking the number of comments for each weibo according to "weibo_comment" column.
- Duplicates which are exact matches of both the "weibo_id" and "weibo_text" columns, i.e. weibos collected repeatedly across keywords. We only keep one of these repeated instances.

This led to a final set of 3,856 weibos, along with their associated weibo IDs which we use in the next step to retrieve comments.

Step III: Extract Comment Data

In order to extract comments for the collected weibos, we utilise their weibo ID. This enabled us to collect textual content and metadata of weibos, including user profiles of commenters. This led to the collection of 31,677 comments for the 3,856 weibos.

Step IV: Process Comment Data

For processing the comments collected in the previous step, we remove comments matching at least one of the following criteria:

- Remove duplicate comment texts, keeping only one instance. This is caused by users who copy and paste the same comment repeatedly.
- Remove short comments with commonly identified patterns – fixed tokens on Sina Weibo, e.g., comments solely containing the word "转发" (repost), "回复" (reply) or "举报" (report).
- Remove the remaining short comments (length of less than 5 characters).
- Remove comments without any Chinese characters.

Given that users occasionally reply by splitting their texts into multiple comments, we aggregate them. When we find multiple comments from the same user in close temporal proximity, we automatically aggregate them into a single comment.

Finally, we convert all the comments from traditional Chinese to simplified Chinese, which helps ensure consistency while keeping the same information. We use the Python package chinese_converter[20] to achieve this.

This led to a final set of 8,969 comments linked to 1,527 weibos, whose statistics are shown in Table 4.4. The final aim of our sexist data collection lies in the retrieval of

[20] https://pypi.org/project/chinese-converter/

these comments, which are the ones that we annotate and make up the final dataset. The weibos are solely considered to support the annotation process and, if desired, for context-based analysis of comments.

4.2.3 Ethics of Data Collection

Due to limitations on the number of weibos that can be crawled at one time and the continuous changes of the Sina Weibo API, we directly obtain the weibo contents via web scraping by using a Python script. Hence, we carefully consider the ethical implications behind the collected data. Posts and comments collected in this dataset are in the public domain and web scraping has been done only for research purposes. Hence, we ensure that no ethics approval is needed for this study [192] and that the collected dataset follows acceptable ethical practices by adhering to the following:

- Our dataset does not present any personally identifiable information, as we have anonymised all user names in the dataset, including any user names mentioned in the posts (replaced by the special token <username>)
- Our dataset does not include any private messages between users, and there was no interaction between Weibo users and researchers.
- We rely on publicly available data and carefully collect the data into multiple steps to avoid overloading Sina Weibo servers.
- The Sina Weibo server is publicly accessible.

4.2.4 Limitations

Regarding our keyword-driven methods for data collection, although these keywords provide a comprehensive framework for capturing discussions related to sexism, their

selection can introduce several potential biases in our final dataset. These keywords may attract certain types of conversations and exclude others, such as predominantly capturing feminism-related expressions with the majority of weibos collected by the term "feminism". And the context in which these keywords are used can vary significantly based on cultural and linguistic factors, potentially reflecting specific cultural attitudes more than others. Keywords tied to specific events, such as "MeToo movement", might result in temporal spikes in data collection, skewing the dataset towards particular periods or events rather than providing a consistent representation over time. Additionally, focusing solely on these keywords might overlook other relevant discussions about sexism that do not explicitly mention these terms. By acknowledging these potential biases, we aim to approach our data collection and analysis with a critical perspective, ensuring that we interpret the findings within the context of these limitations.

4.3 Data Annotation

During the annotation process of our SWSR dataset, we perform three annotation tasks as follows:

1. **Sexism Identification:** whether a text is sexist, as a binary annotation task determining if a comment is sexist (1) or non-sexist (0). Where a comment is deemed sexist, we also perform two additional annotations:

2. **Sexism Category:** We define four categories of sexism, namely stereotype based on appearance (SA), stereotype based on cultural background (SCB), microaggression (MA) and sexual offence (SO).

3. **Target Type:** individual (I) or generic (G).

4.3.1 Annotation Preparation

In order to reliably identify sexism as well as its corresponding categories and targets, we provide initial annotation guidelines for all three tasks. The annotation guidelines for sexism identification are based on [14, 60], and guidelines for the sexism category and the target type are adapted from [23, 46, 61, 64]. Guidelines were iteratively developed through collective annotation of a small sample of 100 comments by a broader set of five annotators. These annotators met and discussed disagreements between them, which led to revised guidelines.

In most cases, we find that our disagreement with annotation task I was mainly caused by the lack of sufficient context when identifying sexist content. For example, one annotator marks the text 它们的大脑平滑到可以在上面溜冰，真的不是一个物种啊[21] as not sexism because there is no sexist content towards women. But when we check the original Weibo text, we find that "they" in this text is intended by its author to mean "some stupid women who insult men for more benefits". So it should be marked as sexism with consideration of the context. Another common case of disagreement is the misunderstanding of specific words related to sexism. These words commonly appear in sexist text but are not common in general speech. Some annotators did not realise that 婚驴 (marriage donkey) is an offensive word specifically towards women. People who use this word have the intention to depict the image of "women who are as stupid as donkeys in marriage, deprived of a lot of benefits, but still enjoy silly happiness". Discussions following these agreements led to revisions in the guidelines and improvements in subsequent rounds of annotations. In addition, for the annotation task II determining the sexism category, there were disagreements caused by occasional overlaps in the interpretations of the different labels, which were resolved and led to revision of the guidelines. Annotation III consisting of determining

[21] Translation: Their brains are so smooth that they can skate on them. We are really not the same species

the target type was more straightforward as it is easier to label.

In what follows, we reproduce the initial guidelines used for the three annotation tasks, which enable annotators to have a better understanding of sexist issues for three annotation tasks and to a large extent improve the final score of inner-annotator agreement.

4.3.2 Annotation Guidelines

Given the difficulty of identifying sexist behaviours, we carefully crafted guidelines for the three annotation tasks based on the insights from the above annotation testing: sexism identification, sexism category and target category, along with examples of annotations by sexism category and target category shown in Table 4.2.

Annotation I: Sexism Identification

A comment is considered sexist if it belongs to at least one of the following categories:

- explicitly attacks or insults gender groups or individuals using sexist language.
- incites gender-based violence or promotes sexist hatred but does not directly use a sexual abusive language.
- abuses those who attack or have negative attitudes towards a gender group.
- shows support for problematic incidents or intentions of sexual assault, sexual orientation and sexually harassment.
- negatively stereotypes gender groups by describing physical appeal, oversimplifying images or expressing the superiority of men over women.
- expresses underlying gender bias sarcastically or tacitly.

The rest of the texts are considered non-sexist. This includes neutral descriptions or testimonies of sex-related events or phenomena.

Annotation II: Sexism Category

Each of the comments marked as sexist in the first task needs to be classified into one of the following, determining the sexism category of the comment:

- *Stereotype based on Appearance (SA):* describes physical appeal, oversimplifies image, or makes comparisons with narrow/vulgar standards towards a gender group.
- *Stereotype based on Cultural Background (SCB):* expresses opinions indicating the superiority of men over women and emphasises gender inequality under the concept of a patriarchal society.
- *Microaggression (MA):* intentionally or unintentionally expresses hostile, derogatory or negative attitudes or remarks against gender groups or individuals.
- *Sexual Offense (SO):* incites sexual-related behaviour or attitude against women, such as sexual harassment, sexual assault, rape and violence.

Annotation III: Target Category

Each of the comments marked as sexist in the first task needs to have the type of target identified, which can be one of the following two:

- *Individual (I):* a post with sexist content addressing a specific person.
- *Generic (G):* a post with sexist content addressing a broader group (such as a gender-based group of people).

Table 4.2: Examples of sexism categories and target types in the dataset.

Example	Translation	Sexism Category	Target
前任的漂亮更清纯甜美一看就是正经人，现在这位一看就很肉的感觉	His ex looks more innocent and beautiful, like a decent person. But the appearance of his current girlfriend makes me a higher libido.	SA	I
还是让女性做些带孩子，鼓励丈夫的工作！	We should let women do more housework, and encourage their husbands' work!	SCB	G
关键是有些女生还没子宫道德，结了婚脑子里自动长了个⊠	The point is that some girls have no uterine morals. There is a dick in their head after they get married.	MA	G
你全家女性送来给我搞一搞，我戴套，保证安全	Send your family's women to me to fuck them, I will wear a condom to ensure safety	SO	I

4.3.3 Annotator Agreement

All three annotations were performed independently by three annotators, all of them PhD students, including two females and one male. We use the open source text annotation tool *doccano*[22] to facilitate the annotation work and to enable independent annotation effectively by three annotators.

We report inter-annotator agreement rates for the three annotators by using Cohen's kappa as a metric [193]. The inter-annotator agreement of our annotation task I is overall 82.3% (71.8% for the sexist class and 96.1% for non-sexist). For annotation tasks II and III, the inner-annotator agreements reach 76.8% and 85.5% respectively. All these agreement rates can be deemed substantial agreements between the three annotators. Examples of annotations by sexism category and target category are shown in Table 4.2.

[22] https://github.com/doccano/doccano

4.4 Lexicon Collection

We build a large sexism and hate lexicon SexHateLex by aggregating and expanding existing resources, which is a combination of

- profane words and slang,
- sexual abusive words and slang, and
- sexism-related people, websites and events.

SexHateLex is built by integrating four existing lexicons, and augmented by adding typos and synonyms based on integrated sexual-related abusive terms. We aggregate the following lexical resources:

- *Chinese Profanity in Wikipedia:*[23] Wikipedia provides a list of Chinese profane words linked to sex, race and sexual orientation. For our purposes, we chose the 599 terms for sex and sexual orientation.
- *HateBase:*[24] HateBase is the world's largest structured repository of regionalised multilingual hate speech corpora in the field of religion, gender, nationality, ethnicity, etc. We collected 29 Chinese terms from HateBase.
- Tocp *dataset:*[25] Ntou Chinese Profanity (Tocp) is the largest Chinese profanity dataset including 16,450 sentences [194]. All profane words and corresponding locations in each sentence have been labelled in this dataset. A total of 1,014 profane words were extracted.

[23] https://en.wikipedia.org/wiki/Mandarin_Chinese_profanity#Sex
[24] https://hatebase.org/
[25] http://nlp.cse.ntou.edu.tw/resources/TOCP/

- *Sexy Lexicon:*[26] The repository funNLP provides massive resources to support research in Chinese NLP, one of which is a sexy lexical list in the category of sensitive term datasets. We collected 1,240 terms from this list.

After integrating terms from all the resources above, we get a total of 2,109 terms. Then we combine typo words that users make spell mistakes in the text based on a spell checking method in the "aion" python package,[27] and add the top 5 similar words to each word in the collected lexical list. FASTTEXT word embeddings[28] are leveraged for this step, followed by cleaning all duplicate and incorrect terms. This leads to the final SexHateLex lexicon with 3,016 terms.

4.5 Data Description

We describe the resulting dataset by first presenting the dataset structure and then providing descriptive statistics of the dataset.

4.5.1 Dataset Structure

Table 4.3: Description of features in the weibo and comment datasets.

Table	Feature
SexWeibo	weibo_id, weibo_text, keyword, user_gender, user_location, user_follower, user_following, weibo_like, weibo_repost, weibo_comment, weibo_date
SexComment	weibo_id, comment_text, gender, location, like, date, label, category, target

The SWSR dataset is organised in two files: *SexWeibo.csv* (SexWeibo) and *SexComment.csv* (SexComment), containing weibos (posts) and comments (replies) respectively. Contents in these two files can be linked through the *weibo_id*. We list

[26] https://github.com/fighting41love/funNLP/tree/master/data
[27] https://github.com/makcedward/nlp/tree/master/aion
[28] https://fasttext.cc/

all features in *SexWeibo.csv* and *SexComment.csv* files in Table 4.3 (see more details in B). Considering user privacy, all user names in this dataset are anonymous with a special token <username>.

4.5.2 Dataset Statistics

Table 4.4: Statistics of the dataset.

	All	Sexist	Non-Sexist
All	8969	3093 (34.5%)	5876 (65.5%)
Average length per comment	71.45	90.34	61.51
Number of comment per weibo	5.87	3.77	4.69

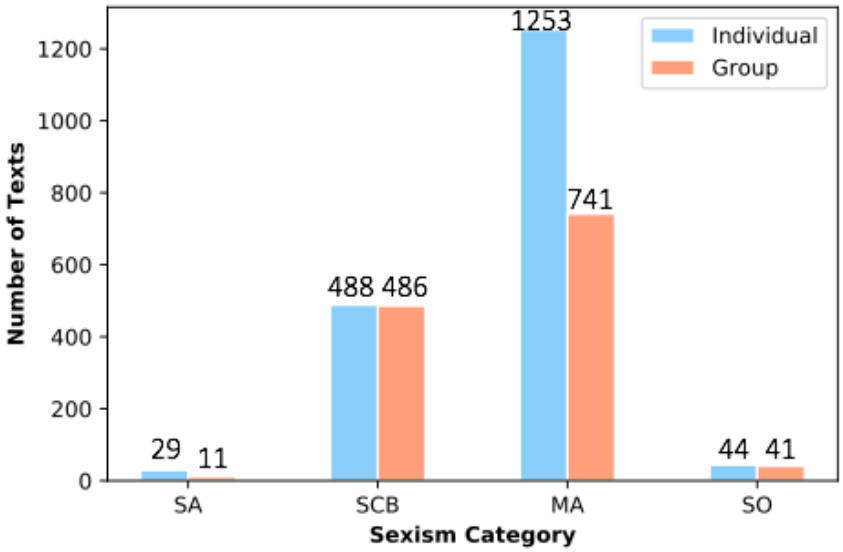

Figure 4.3: Distribution of sexism categories and target types in the dataset.

The resulting 8,969 comments are associated with 1,527 weibos. Table 4.4 shows the statistics of the dataset in terms of the distribution of sexist comments, comment length and number of comments per weibo. We can see that the majority of comments are non-sexist, with nearly twice as many as sexist comments.

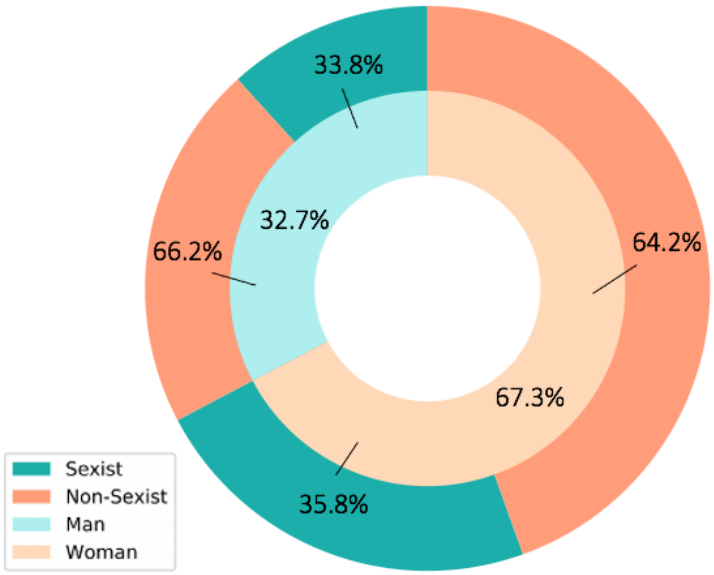

Figure 4.4: Distribution of user gender across two classes in the dataset.

Figure 4.3 depicts the distribution of the sexism category and target type in sexist comments. More than half of the sexist comments are MA, and SCB also takes a large proportion in the sexist class. Besides, the number of comments towards individuals nearly doubles those towards groups, where sexist texts in the MA category are more frequently abusive towards individuals.

Textual Distribution

We compute the average lengths (in a number of characters) of comments in each category. We see big differences in Table 4.4 showing that the average length of a sexist comment is 50% bigger than the length of a non-sexist comment. Furthermore, Table 4.4 presents the average number of comments for each weibo. We can see that the number of comments per weibo for both sexist and non-sexist classes is less than that for all data, which is because that one weibo might contain multiple comments in different classes. Hence, the sum of weibo counts for two classes can be larger than

the overall number of weibos.

Gender Distribution

Figure 4.4 shows that 32.7% of posts are made by male users and 67.3% by female users. Among these, 33.8% of male posts and 35.8% of female posts are sexist, indicating a similar tendency towards posting sexist content across genders, with a slightly higher rate among women. Sexist texts may differ by gender – men often use more direct and aggressive language, while women may express more subtle forms of internalised misogyny and benevolent sexism [18, 19]. And the higher engagement of female users in sexist posts suggests that the discussions around specific keywords might attract more female participation, possibly due to greater individual relevance or impact of these topics [24]. Additionally, the selected keywords may have a certain impact on constructing the dataset. These terms are biased and more prominently reflect specific aspects of sexism. This potentially leads to over-representation or under-representation of certain types of sexist behaviours, thus affecting the overall analysis and downstream tasks [32, 195]. For this situation, future analysis should explore the correlation between topics and male/female user behaviours to ensure a comprehensive understanding of gender dynamics in online sexist discussions.

Word Frequency Distribution

We normalise the data by removing stop words, special markers such as "转发" (Repost), user names, and punctuation marks. Then we select the list of 12 words with the highest frequency in the comments as well as the top 12 words from the SexHateLex lexicon which are most frequent in the comments. We find that the terms frequently occurring in each class differ significantly (see Table 4.5). The most frequent tokens in the lexicon present negative emotional attitudes while those in the comments are mostly neutral words related to gender topics.

Table 4.5: Description of the 12 most frequent terms in the dataset (DataTerm) and in the lexicon (LexTerm). [尸吊] is a sensitive character which cannot be found in the Latex package. The table presents the character by dividing it into two parts, which can be easily understood in Chinese. PCT denotes the percentage of each term.

DataTerm	Translation	PCT	LexTerm	Translation	PCT
女权	feminism	29.84%	骂	curse	7.45%
女性	women	25.20%	死	die	2.89%
不是	not	19.04%	搞	flirt	2.75%
男人	man	11.92%	女拳	negative feminism	2.20%
孩子	children	8.78%	歧视	discrimination	2.04%
骂	curse	7.45%	驴	donkey	1.88%
男权	patriarchal	6.14%	[尸吊]	dick	1.78%
极端	extreme	5.90%	逼	pussy	1.45%
结婚	marry	5.26%	强奸	rape	1.44%
姓	surname	5.26%	狗	dog(similar use as pig)	1.23%
权利	right	3.89%	干	fuck	1.08%
平等	equality	3.73%	蛆	maggot	0.89%

4.6 Preliminary Experiments: Sexism Detection

To assess the difficulty of computationally detecting sexist comments in SWSR and to provide benchmark experimental results, we conduct both coarse-grained and fine-grained sexism detection experiments, evaluating different features and models. Our experiments are designed in three steps:

1. Sexism identification (Binary): weibo contents are classified as either sexist or non-sexist.

2. Sexism category classification (Multi-class): texts are classified into one of five categories: stereotype based on appearance (SA), stereotype based on cultural background (SCB), microaggression (MA), sexual offence (SO), or non-sexist.

3. Target classification (Multi-class): texts are classified into either generic, individual, or non-sexist.

4.6.1 Models

For the three experimental steps, we test various models. As context-based models we utilise different BERT-based models [33] based on transformers. We use three different BERT-based models: (1) BERT, (2) BERT with whole word mask (BERT-wwm), and (3) Robustly optimised BERT approach (RoBERTa) [196]. Besides, we adopt three different baselines using combinations of unigrams to trigrams as features: (1) Logistic Regression (LR), (2) Support Vector Machine (SVM), and (3) a character-level LR (char-LR). We also test two content-based models, Convolutional Neural Network (CNN) and character-level CNN (char-CNN) [197] with FASTTEXT word embeddings.

In addition, for the experimental step 1, we test all the models above with and without lexical words from the SexHateLex lexicon, to show its impact on the task. We first count the occurrence of each word, and then convert the count vector from the count frequency to Term Frequency–Inverse Document Frequency (TF-IDF) [198], indicating how significant a category is to a text in the corpus. Finally, we concatenate the TF-IDF lexical vector with textual embeddings. For BERT-based models, we concatenate lexical embeddings with the output of BERT, and then feed them into a feedforward layer for final classification.

4.6.2 Experiment Settings

Given that the SWSR dataset is not balanced, especially in the category classification task, we randomly split the comment data into 90% for training and 10% for testing using stratified sampling. Class distribution in the training set includes 34.7% sexist texts and 65.3% non-sexist texts. We perform cross-validation experiments on the training data to fine-tune model hyperparameters, choosing the best models for the final experiments. We report global macro F1 and accuracy scores for the three tasks, as well as F1 scores specific to each class for experimental step 1 and weighted F1 scores for steps 2 and 3.

4.6.3 Experiment Results

Table 4.6: Sexism detection performance. F1-Sex and F1-Not denote F1 scores respectively for binary labels of sexist or non-sexist. mF1 denotes macro F1 score and Acc denotes accuracy score.

Model	Original Feature				+Lexicon			
	F1-Sex	**F1-Not**	**mF1**	**Acc**	**F1-Sex**	**F1-Not**	**mF1**	**Acc**
LR + ngram	0.624	0.849	0.737	0.785	0.616	0.846	0.731	0.780
char-LR + ngram	0.640	0.852	0.746	0.790	0.646	**0.858**	0.752	0.797
SVM + ngram	0.633	0.844	0.739	0.781	0.640	0.842	0.741	0.786
CNN + ft	0.669	0.828	0.749	0.774	0.654	0.844	0.749	0.785
char-CNN + ft	0.660	0.845	0.753	0.787	0.654	0.850	0.752	0.790
BERT	**0.694**	**0.858**	**0.776**	**0.806**	0.661	0.844	0.752	0.786
BERT-wwm	0.678	0.846	0.762	0.792	0.699	0.851	0.775	0.800
RoBERTa	0.685	0.844	0.764	0.792	**0.707**	0.853	**0.780**	**0.804**

From the results in Table 4.6, we see that content-based models (CNN) outperform linguistic ones (LR and SVM) in both word level and character level while context-based models (BERT) perform best. Character-level models (e.g., char-LR and char-CNN) show better performance than word-level models (e.g., LR and CNN), proving them more suitable for a language like Chinese with no space between words. When we incorporate lexical features, most models lead to slight improvements of 0.5-1% in F1 score (except for LR and BERT models), showing the potential of SexHateLex in improving performance, particularly with the best-performing model RoBERTa. We also observe an overall tendency for achieving 15-23% better prediction in the non-sexist category, highlighting the challenge of detecting sexist comments.

Regarding the category classification task, the results in Table 4.7 show a different scenario. The best-performing model is RoBERTa, with the highest weighted and F1 scores, but all three BERT-based models have better performance than others. For the third task, the results in Table 4.7 show that all the models achieve a competitive performance without a large margin, while RoBERTa performs best across other models. Besides, it can be observed that macro F1 scores for both tasks 2 and 3 show an average lower than weighted F1 scores, which indicates a potential impact

Table 4.7: Results for the sexism category and target classification tasks. mF1 denotes macro F1 score and wF1 denotes weighted F1 score. Acc denotes accuracy score.

Model	Category classification			Target classification		
	wF1	**mF1**	**Acc**	**wF1**	**mF1**	**Acc**
LR + ngram	0.628	0.310	0.611	0.663	0.447	0.719
char-LR + ngram	0.648	0.316	0.646	0.657	0.428	0.721
SVM + ngram	0.647	0.320	0.692	0.661	0.446	0.707
CNN + ft	0.711	0.335	0.716	0.668	0.447	0.711
char-CNN + ft	0.722	0.347	0.730	0.670	0.448	0.714
BERT	0.732	0.355	**0.736**	0.678	0.457	0.713
BERT-wwm	0.732	0.354	**0.736**	0.682	0.462	0.720
RoBERTa	**0.734**	**0.360**	0.732	**0.687**	**0.467**	**0.727**

of the imbalanced nature of the data among the finer-grained classes. More sampling methods are supposed to be considered before training.

4.6.4 Error Analysis

Table 4.8: Error analysis for misclassified examples. TL denotes true label and PL denotes predicted label.

Error Type	Example	Translation	TL	PL
(1)	如果她自己够优秀就不会在网络上怨天尤人了	If she is excellent enough, she won't blame others on the Internet	1	0
(2)	你这种金针菇明码标价了也只会烂在货架上	Enoki mushrooms like yours will only rot on the shelf even if they are clearly marked	1	0
(3)	田园女权，女拳师，极端女权，是我是我都是我	Pastoral feminist, female boxer, extreme feminist, it's all me	0	1

We look at frequent errors across misclassified instances generated from SVM, CNN and BERT, three typical models selected from three types of models we used in the experiment step 1 (see Table 7.5 for examples). Several typical errors appeared in the experiments are summarised below:

(a) Implicit sexism. Errors in those posts lack explicit sexist expression or context, and the most frequent reason misclassified texts in (a) is caused by sarcastic expressions. Sarcasm seems to be a suitable way for expressing contempt and subtly offending individuals, which modifies the perception of the message, hindering the

correct detection of sexism by automatic systems [199]. Example (1) is a sarcastic comment that criticises women who are not successful but insults those people who uphold gender equality. It is difficult to identify sexism when there is no explicit presence of abusive language. Another problem is that the model cannot pick up words with a specific meaning related to gender.

(b) Lack of prior information. It demonstrates that the model cannot identify those contents referring to sexism-related events, people or words/phrases with special meanings as it does not possess prior knowledge. In example (2), 金针菇 (enoki mushroom) is a very harmful word specifically towards men associated with some physical characteristics but cannot be directly identified by the model.

(c) Overuse of sexist words. It indicates that sexist words might be overused in one text, leading to the over-dependence of the model on these words, while sexist targets in posts are confounding and hard to identify. We can see from example (3) that the model can easily identify a text with many sexist words as a sexist text even if there is no specific targeted individual or group attacked by someone.

4.7 Research Applications

The Swsr dataset and the SexHateLex lexicon provide resources for furthering research in a new language in the growing research problem of sexist language. We discuss potential areas of research.

4.7.1 User-based Sexism Detection

As sexism-related speech belongs to user-generated content online, some investigations are conducted to find out the potential influence of user characteristics like gender and location on sexism detection [14]. User metadata in Swsr, such as gender, location and number of followings, can enable researchers to explore possible correlations

between gender-based hateful content and user profiles, furthering user-based studies in the area of sexism detection.

4.7.2 Explainable Sexism Detection

Providing explanations can make model outputs more convincing and understandable [200, 201]. We provide our dataset with two basic classes to show which text is sexist or not, with fine-grained labels to support further detection. Besides, we offer a lexicon composed of abusive words to support the detection of offensive content with sexism-specific features.

4.7.3 Multi-lingual and Cross-lingual Sexism Detection

While most approaches to sexism detection have been proposed for English, other studies have been investigated to deal with this task in other languages such as Spanish, Italian, and Indian, thanks to recent shared tasks [46, 61, 110]. More research is needed in other languages, including Chinese, both in multilingual settings, i.e. proposing models that deal with multiple languages, and cross-lingual settings, i.e. leveraging data in a resource-rich language like English for application in lesser-resourced languages such as Chinese. Our dataset compensates for the lack of sexist speech in Chinese, thereby facilitating the development of sexism identification research in multi-lingual and cross-lingual settings.

4.7.4 Cross-domain Hate Speech Detection

With the prevalence of identifying hate speech online, some studies concentrate on detecting specific types of hate speech, such as racism or sexism. These differences across types of hate speech make it more challenging to generalise hate speech detection models. Cross-domain detection of hate speech thereby has been a topic of interest to identify common features between distinct hate speech domains, achieving knowledge

transfer and model generalisation. Our dataset provides gender-related hateful texts with corresponding topic-related keywords, which could enhance research on sexism and facilitate potential research of cross-domain detection in this and other types of hate speech, particularly if additional Chinese hate speech datasets are released.

4.7.5 Other Applications

While most existing research on sexism detection focuses on detecting the text to binary classes (sexist or not), our dataset enables investigation of additional, finer-grained perspectives of sexism, thanks to three types of labels provided. Categorising sexism by type as well as identifying the type of targets enable furthering research in sexism detection beyond the widely-studied binary classification task.

4.8 Conclusion

In this chapter, we release a comprehensive sexism dataset SWSR along with a large lexicon SexHateLex, to facilitate research on online gender-based speech in Chinese. To the best of our knowledge, this is the first sexism dataset in Chinese. The dataset provides both weibo and comment texts, as well as three types of labels, namely sexist or not, sexism category and target type. The dataset contains two files for *SexComment* and *SexWeibo*, containing sexist comments, original weibos enabling contextual analysis, and anonymised user metadata. We further conduct exploratory analyses of the dataset. Different types of sexism detection approaches are also evaluated on *SexComment*. We experiment with baseline models for sexism detection, which provides a benchmark for further experimentation. We expect our dataset to enable further research in Chinese sexism detection, including a set of possible directions.

5

Multi-Channel Joint Learning for Cross-Lingual Sexism Detection

In this chapter, we investigate the cross-lingual sexism detection task across three languages: English, Spanish and Italian, and introduce the first approach to cross-lingual sexism detection that incorporates capsule networks. We propose a cross-lingual capsule network learning model, a multi-channel architecture coupled with extra domain-specific lexical semantics (called CCNL-EX), and it yields state-of-the-art (SOTA) performance for all six language pairs under study compared with ten baselines.

The chapter is organised as follows. We introduce the architecture of our proposed CCNL-EX model in Section 5.2. In Section 5.3, we describe our experiment settings between CCNL-EX and different SOTA baselines. Then we analyse experimental results and discuss classification errors in Section 5.4.

5.1 Introduction

To further research in broadening the generalisability and suitability of models across languages for sexism detection [75], we incorporate capsule networks and hate-related lexicons to further boost cross-lingual performance.

Capsule Network is a clustering-like method proposed by Sabour et al. [202]. They replace scalar-output feature detectors of Convolutional Neural Network (CNN) with vector-output capsules to learn spatial relationships of entities via dynamic routing, improving representations against CNN. Hinton et al. [203] then propose a new iterative routing based on the Expectation–Maximisation (EM) algorithm, which shows potential in image analysis [202] and is soon applied to Natural Language Processing (NLP) research [204]. Its use on hate speech and sexism detection is however limited [205, 206]. Srivastava et al. [205] put forward a capsule-based architecture for aggressive language classification, and further incorporate multi-dimensional capsules for the same task [206]. Capsule network has not been considered in cross-lingual sexism detection so far. Hence, we contribute to gaps in both lines of research bringing

together cross-lingual sexism detection and capsule network.

We propose a Cross-lingual Capsule Network Learning model with Extra lexical semantics specifically for sexism (CCNL-EX), whose two-parallel framework enriches input information in both source and target languages. The model can be applied to new languages lacking annotated training data [40, 207] and is able to capture spatial positional relationships between words to improve the generalisability of capsules. It can also be exploited to broaden the detection capacity of linguistically diverse genres such as social media.

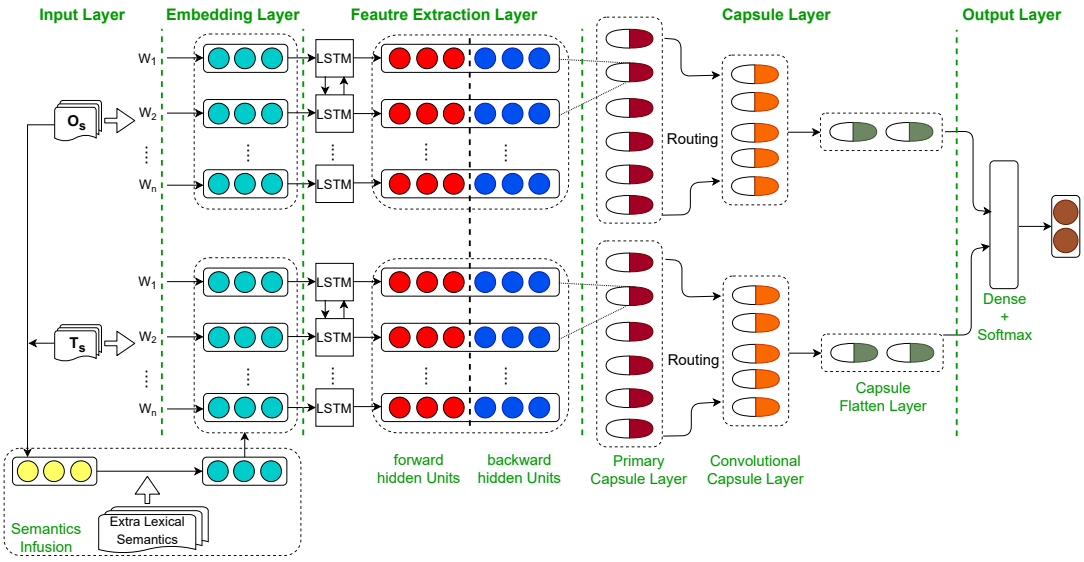

Figure 5.1: The architecture of CCNL-EX.

5.2 Methodology: CCNL-Ex

5.2.1 Model Architecture

Inspired by Capsule Networks built by Sabour et al. [202], we propose a cross-lingual capsule network learning model with multilingual word embeddings integrated lexical semantics (CCNL-EX). It is composed of six layers (see Figure 5.1):

Input Layer

CCNL-EX has two parallel capsule-based architectures for bilingual input training data – O_s in the source language and the parallel translated T_s in the target language.

Embedding Layer

The input data is the sequence of texts and each text consists of a series of words. The input representation is a weight matrix $X \in \mathbb{R}^{e \times V}$ for e-dimensional vector of words and vocabulary size of V, fine-tuned by absorbing extra hate-related lexical semantic information (see Section 5.2.2).

Feature Extraction Layer

In each aligned network, we use a Bidirectional Long Short Term Memory (BiLSTM) network [208] as the feature extractor to get contextual relationships from local features. The output of BiLSTM is $h_t = [h_t^f, h_t^b] \in \mathbb{R}^{(2 \times k)}$, combined by forward feature h_t^f and backward feature h_t^b with k units.

Capsule Layer

The capsule layer consists of a primary capsule layer and a convolutional capsule layer. The primary capsule layer extracts instantiation parameters to represent spatial position relationships between features, like the local order of words and their semantics [204]. Suppose $W \in \mathbb{R}^{(2 \times k) \times d}$ is a shared matrix, where d is the dimensionality of capsules. For the hidden feature h_t, we create each capsule $p_i \in \mathbb{R}^d$:

$$p_i = g(W^T h_t + b) \tag{5.1}$$

where g is a non-linear squash function to compress the vector length between 0 and 1:

$$g(s) = \frac{\|s\|^2}{1 + \|s\|^2} \frac{s}{\|s\|} \tag{5.2}$$

The convolutional capsule layer is connected to capsules in the primary capsule layer. Some primitive routing algorithms, like max pooling in CNN, only capture features to show whether it exists in a certain position or not, missing more spatial relationships [204]. The connection weight is learnt by a dynamic routing, which can reduce the loss to make the capsule network more formative and effective, and attach less significance to unrelated or useless content, such as stop words [205]. The process of dynamic routing between the primary capsule u_i and the convolutional capsule v_j is as below:

$$c_{j|i} = softmax(b_{j|i}) \tag{5.3}$$

$$v_j = g(\sum_i c_{j|i} u_{j|i}) \tag{5.4}$$

$$b_{j|i} = b_{j|i} + u_{j|i} \cdot v_j \tag{5.5}$$

where $b_{j|i}$ denotes the connection weight between capsules. After the routing, all output capsules are flattened.

Output Layer

The final representations from the two parallel architectures are concatenated, using a softmax function to obtain the label probability.

5.2.2 Lexical Semantic Knowledge Infusion

We fine-tune pre-trained word embeddings by infusing domain-specific lexical semantic knowledge, aiming to obtain domain-aware word representations and enhance the model capacity of identifying hate-related content. More specifically, we firstly retrieve the five most relevant semantic words from SenticNet [209] for each lexical

word, and utilise FastText embedding model [128] to generate the five most similar words for each Out-Of-Vocabulary (Oov) word. Then we apply the similarity learning method proposed by Faruqui et al. [38] to integrate lexicon-derived semantic information into pre-trained word embeddings by minimising distances between a word and its semantically related words.

5.3 Experiments

We investigate cross-lingual sexism detection as a binary classification task in three different languages –English (EN), Italian (IT) and Spanish (ES)– and all six possible language pairs involving them: ES→EN, EN→ES, IT→EN, EN→IT, ES→IT, and IT→ES.

5.3.1 Datasets

We use gender-based sexism datasets from the Automatic Misogyny Identification (AMI) tasks held at the Evalita 2018[29] and IberEval 2018[30] evaluation campaigns. These datasets provided by AMI@Evalita and AMI@IberEval are extracted from the Twitter platform, and constructed under the same annotation scheme for binary labels: misogynistic and non-misogynistic. AMI@Evalita datasets present texts in English and Italian [46], while the AMI@Ib-erEval ones are in Spanish and English [61]. We utilise English data only from AMI@Evalita to make data size balanced among three languages, as well as for consistency with previous research [40] to enable direct comparison. Given the well-divided training and test sets for each language, we further randomly select 20% of the training set as the validation set for the model fine-tuning process, and finally utilise the whole training set to evaluate the model capacity on the test set. More details of datasets can be seen in Table 5.1. We create

[29] https://amievalita2018.wordpress.com/data/
[30] https://amiibereval2018.wordpress.com/important-dates/data/

parallel corpora for separate datasets by directly using Google Translate[31] to translate all data between source and target languages.

Table 5.1: Distribution of train, validation and test sets, misogynistic text rate (MTR) in source training and test sets, data sources for three languages.

Language	English (EN)	Spanish (ES)	Italian (IT)
Train	3200	2646	3200
Validation	800	661	800
Test	1000	831	1000
MTR$_{train}$ (%)	44.6	49.9	45.7
MTR$_{test}$ (%)	46.0	49.9	50.9
Source	Evalita 2018	IberEval 2018	Evalita 2018

5.3.2 Multilingual Lexicons

To further assess model performance, we integrate two multilingual domain-related lexicons as extended knowledge into embeddings to explore the possibility of fine-tuning word embeddings and investigate their potential to further boost performance:

- *HurtLex:*[32] It is a multilingual hate speech lexicon, containing offensive, aggressive, and hateful words or phrases in over 50 languages and 17 categories. We obtain 6,287 words for English, 3,565 for Spanish and 4,286 for Italian from it.

- *Multilingual Sentiment Lexicon:*[33] Since hate speech and sexist content often express more negative sentiments [210], we utilise a sentiment lexicon, which consists of positive and negative words in 136 languages, and provides 2,955 negative words for English, 2,720 for Spanish and 2,893 for Italian [211].

[31] https://translate.google.co.uk/

[32] http://hatespeech.di.unito.it/resources.html

[33] https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

5.3.3 Baselines

We compare both CCNL and CCNL-EX (with lexicons infused) with ten baselines, including SVM with unigrams features, CNN, BiLSTM and CapsNet with monolingual FASTTEXT embeddings of translated target data, multilingual embeddings MUSE and LASER fed to a 2-layer feedforward neural network, the SOTA cross-lingual models mBERT and XLM-R covered by a 2-layer feedforward classifier on the output layer, and hate-specific cross-lingual model JL-HL with two inputs proposed by Pamungkas and Patti [40]. All baselines are described as follows:

Majority: The majority classifier always predicts the most frequent class in the training set. MTR_{train} values for three languages are less than 50%, which means the majority class is non-misogynistic.

SVM: Support Vector Machine (SVM) aims to determine the best decision boundary between vectors that belong to a given category or not [212].

CNN: Convolutional Neural Network (CNN) consists of one convolutional layer and one max pooling layer to capture local textual features [197].

BiLSTM: Bidirectional Long Short-Term Memory (BiLSTM) is composed of forward/backward recurrent neural networks to extract long-term dependencies of a text [213].

CapsNet: A single capsule network (CapsNet) [202] uses a convolutional layer to extract n-gram features.

LASER: Language-Agnostic SEntence Representations (LASER) aims to calculate and use joint multilingual sentence embeddings across 93 languages [131].

MUSE: Multilingual Unsupervised and Supervised Embeddings (MUSE) builds bilingual dictionaries and aligns monolingual word embedding spaces without supervision [81].

mBERT: Multilingual BERT[34] (mBERT) is a variant of BERT [33] that was trained

[34] https://github.com/google-research/bert/blob/master/multilingual.md

on 104 languages of Wikipedia.

XLM-R: XLM-RoBERTa (XLM-R) is a scaled cross-lingual sentence encoder across 100 languages from Common Crawl [36].

JL-HL: A joint-learning cross-lingual model proposed by Pamungkas and Patti [40], a hybrid approach with LSTM architectures which concatenates multilingual lexical features from HurtLex [73]. The source data and translated target data are fed to two parallels separately.

5.3.4 Experiment Settings

For training our model, we use FASTTEXT embeddings of dimension 300 trained on the Common Crawl and Wikipedia [128]. We use 128 units for forward and backward LSTM (256 units in total) and 50 units in the hidden layer for the feedforward classifier. For capsule networks, we use 10 capsules of dimension 16 and the number of dynamic routing is 5. We use Adam optimiser with 0.0001 learning rate, and set 0.4 for dropout value and 8 for batch size. The model is coded in Keras 2.2.4 and Tensorflow 1.14. We run experiments on the HPC resources of our university, each experiment taking less than one hour. Due to the imbalanced nature of the label distribution among datasets, macro-averaged F1 score is reported as the evaluation metric for all experiments.

5.4 Results

5.4.1 Model Performance

Results are shown in Table 5.2. CCNL and CCNL-EX differ in that the latter incorporates lexical semantic features. We can observe that CCNL yields better performance than all baseline models for five out of six language pairs, with the exception of ES→EN. CCNL-EX outperforms all ten baselines for all language pairs. These results

substantiate the effectiveness of our model with semantic information, highlighting its generalisation capability across three languages.

Table 5.2: Comparison of CCNL and CCNL-EX over baselines on the six language pairs. The best result is highlighted in **bold** and the second best result underlined.

Model	ES→EN	EN→ES	IT→EN	EN→IT	ES→IT	IT→ES
Majority	0.351	0.334	0.351	0.329	0.329	0.334
SVM	0.620	0.561	0.588	0.227	0.643	0.525
CNN	0.598	0.613	0.592	0.275	0.636	0.607
BiLSTM	0.575	0.608	0.597	0.341	0.498	0.459
CapsNet	0.616	0.559	0.601	0.323	0.555	0.611
LASER	0.552	0.466	0.597	0.374	0.678	0.619
MUSE	0.592	0.491	0.618	0.400	0.717	0.666
mBERT	0.567	0.580	0.568	0.399	0.648	0.618
XLM-R	0.583	0.618	0.597	0.411	0.677	0.613
JL-HL	0.635	0.687	0.605	0.497	0.660	0.637
CCNL	0.624	0.719	0.628	**0.584**	**0.735**	**0.668**
CCNL-EX	**0.651**	**0.729**	**0.629**	0.519	0.736	**0.670**

Among the ten baselines, the best is JL-HL, whose performance is still always below that of CCNL-EX. CCNL achieves absolute improvements ranging 7%-9% over JL-HL model for two language pairs involving Italian: EN→IT, and ES→IT. In addition, the CCNL model has manifested pronounced improvements in terms of separately identifying two classes compared to the majority baseline. We can also observe that CCNL generally achieves a large margin with respect to other baselines like SVM, CNN and BiLSTM (especially for ES→EN, EN→IT and IT→ES), while the baseline MUSE achieves good results for IT→EN and IT→ES. It also highlights the effectiveness of the capsule network as an important component in the cross-lingual model compared with other baselines. Furthermore, we observe that CCNL achieves better performance on all six language pairs when we compare it with CapsNet, LASER, mBERT and XLM-R. Possible reasons are that BiLSTM layers enable the proposed model the capability of extracting contextual information compared to the CNN layer, and CCNL takes the spatial features into consideration by sharing the same weight matrix and learning the positional feature difference in high level via the dynamic routing process.

Compared with CCNL, CCNL-EX performs better for ES→EN and EN→ES, and shows a similar performance in three out of six language pairs, which indicates the effectiveness of integrating semantics based on lexicons. The exception to the trend showing better performance for lexicon-based methods is for the two language pairs that have Italian as the target, namely EN→IT and ES→IT, where the base CCNL model with no lexicons performs best. This is likely due to limitations in the Italian language lexicons, and hence reinforces the need to secure high-quality lexicons if they are to be incorporated.

5.4.2 Comparative Experiments

In order to explore the effect of diverse components in our cross-lingual capsule model on six language-pair tasks, we further implement experiments to assess ablated models compared to our basic framework CCNL and the impact of varying specific components in the feature extraction layer.

Framework Ablation Analysis

We perform an ablation study for CCNL by dropping one of the two parallel architectures (CCNL-non-parallel), removing the LSTM layer (CCNL-non-LSTM) and removing the Capsule Network layer (CCNL-non-CapsNet). As shown in Table 5.3, CCNL outperforms all ablated models, demonstrating the combined benefits of all CCNL components. CCNL noticeably outperforms CCNL-non-parallel on all language pairs, highlighting the importance of the two-parallel framework for extracting local features from both source and target texts. Additionally, we can validate the ability of the BiLSTM network to extract contextual information effectively compared with CCNL-non-LSTM, which highlights the effectiveness of the capsule network compared with CCNL-non-CapsNet.

Table 5.3: Ablative experiment results for CCNL. The best result is highlighted in **bold**.

Model	ES→EN	EN→ES	IT→EN	EN→IT	ES→IT	IT→ES
CCNL-non-parallel	0.522	0.558	0.570	0.513	0.626	0.624
CCNL-non-LSTM	0.373	0.609	0.565	0.406	0.685	0.623
CCNL-non-CapsNet	0.597	0.678	0.613	0.439	0.643	0.622
CCNL	**0.624**	**0.719**	**0.628**	**0.584**	**0.737**	**0.668**

Impact of Feature Extraction Layer

We aim to validate the ability of the BiLSTM network to extract contextual information effectively. We test different feature extraction layers by keeping other components of the CCNL architecture unchanged. We test four flavours of CCNL: CCNL (with LSTM feature extraction layer), CCNL-non-FE (without feature extraction layer), CCNL-CNN (CNN feature extraction instead) and CCNL-GRU (bidirectional Gated Recurrent Unit feature extraction instead). Results in Table 5.4 show that CCNL with the feature extraction layer performs consistently better than those without it, highlighting the importance of extracting local features from the text. Additionally, CCNL also shows improved performance on all tasks when compared with CCNL-CNN, since contextual information plays a significant role in detecting sexism. CCNL noticeably outperforms CCNL-GRU on four out of six language pairs and achieves similar performance (differences below 0.5%) on the other two language pairs (EN→ES and IT→ES). This is likely due to structural similarities of GRU and LSTM, with the additional complexity of LSTM allowing to capture more informative features in some language pairs.

Table 5.4: Results for different feature extraction layer in CCNL. The best result is highlighted in **bold**.

Model	ES→EN	EN→ES	IT→EN	EN→IT	ES→IT	IT→ES
CCNL-non-FE	0.373	0.609	0.565	0.406	0.685	0.623
CCNL-CNN	0.521	0.592	0.577	0.439	0.633	0.622
CCNL-GRU	0.458	**0.722**	0.613	0.411	0.715	**0.671**
CCNL	**0.624**	0.719	**0.628**	**0.584**	**0.737**	0.668

5.4.3 Qualitative Analysis

In this section, we provide a qualitative analysis of the role of semantic understanding in enhancing the detection of hate speech in multilingual scenarios.

5.4.3.1 Integration of Semantic Knowledge

We examine instances from different language pairs to illustrate how semantics help improve the accuracy of our CCNL-EX. Table 5.5 presents examples that are misclassified by CCNL but classified correctly by CCNL-EX after infusing external lexical information.

Table 5.5: Examples for semantic analysis. Translated texts are presented for non-English instances. Ground truth (GT), prediction labels without lexical knowledge (P), and prediction labels with lexical knowledge (P-Ex) are noted – hateful (1) and non-hateful (0).

#	Text	GT	P	P-Ex
(i)	my mother is a slut bitch	1	0	1
(ii)	il mio puttanone preferito look at her Translation: my favorite whore look at her	1	0	1
(iii)	Eres un idiota sin esperanza Translation: You are a hopeless idiot	1	0	1

We can see from example (i) that the CCNL-EX model accurately identifies explicit derogatory terms such as "slut" and "bitch". The term "puttanone" in example (ii), which is hateful in Italian, is correctly recognised. Example (iii) shows that the term "idiota (idiot)" in Spanish and the context of "sin esperanza (hopeless)" make this text offensive.

These terms can be found either in external lexical resources or their translated versions. These examples illustrate the effectiveness of incorporating external semantic resources in detecting hate speech across different languages. Leveraging lexical knowledge enables the model to understand cultural context, implicit hate, and explicit offensive language, thereby enhancing its performance in cross-lingual hate

speech detection.

5.4.3.2 Error Analysis

We inspect frequent errors across misclassifications from the test set by the CCNL-EX model (see Table 5.6 for examples). We summarise the following four main types of errors:

Table 5.6: Examples for error analysis. Translated texts are presented for non-English instances. Ground truth (GT) and prediction (P) labels are noted – hateful (1) and non-hateful (0), along with corresponding error types (ET).

Text	GT	P	ET
Analicemos esto: ¿Si te pones unos shorts así, en la calle, ¿qué esperas que te digan? ¿Acoso? ¿O Provocación... Translation: Let's analyse this: If you wear shorts like this, in the street, what do you expect them to say? Bullying? Or Provocation ...	1	0	a
tranquille ragazze, tranquilli gay, il Butturini c'ha una morosa che un pezzo di figa mostruosa! #TVOI Translation: quiet girls, quiet gays, Butturini has a girl -friend who is a piece of monstrous pussy! #TVOI	0	1	b
@user ben sasse is 100% correct. since 1973, all ive ever heard every two years for elections are hysterical women (all a leftist act) about back-alley abortions. this shit is getting old! i didn't hear one other protest issue being yelled about i	1	0	c
@user ma se la #culona #tedesca che predica #austerit mi soon perso qualcosa Translation: @user but if the #culona #german preaching #austerit I missed something	1	0	d

(a) Implicit hate. Those lacking explicit hateful content or context in the post;

(b) Overuse of hateful words. Hateful words can be overused, leading to the over-dependence of the model on these words, while hate targets in posts are confounding and hard to identify;

(c) Lack of prior information. The model cannot identify those contents referring to hate-related events, people or words/phrases with special meanings as it does not possess prior knowledge;

(d) Erroneous translation. The use of machine translation can lead to translation errors for important words. Some words used in hashtags cannot be easily translated, which might be regarded as OOV words by the model.

5.5 Conclusion

In this chapter, we propose a Cross-lingual Capsule Network Learning model integrating Extra hate-related semantic features (CCNL-EX) for sexism detection. CCNL, the main framework of our model, is composed of two parallel architectures for source and target languages, using BiLSTM to extract contextual features and Capsule Network to capture hierarchically positional relationships. Our model finally leads to SOTA performance for all six language pairs compared with ten competitive baselines. Results show the potential of learning contextual information and spatial relationships of sexist texts.

6

Leveraging Pre-trained Semantics and Lexical Features for Multilingual Sexism Detection

In this chapter, we introduce a novel architecture based on multilingual pre-trained language models (PLMs) for multilingual sexism identification using EXIST datasets, which is made of the last 4 hidden states of XLM-R and a TextCNN with 3 kernels. We also exploit lexical features relying on the use of new and existing lexicons of abusive words, with a special focus on sexist slurs and abusive words targeting women.

The chapter is organised as follows. In Section 6.2 we introduce more details about EXIST task and datasets. Section 6.3 describes the architecture of our proposed model. Then we present experimental results and analysis for multilingual sexism detection in Section 6.4, along with a discussion in Section 6.5.

6.1 Introduction

Online sexism is an increasing concern for those who experience gender-based abuse in social media platforms as it has affected the healthy development of the Internet with negative impacts on society. The EXIST shared task proposes the first task on sEXism Identification in Social neTworks (EXIST) at IberLEF 2021 [214]. It provides a benchmark sexism dataset with Twitter and Gab posts in both English and Spanish, along with a task articulated in two subtasks consisting of sexism detection at different levels of granularity: Subtask 1 Sexism Identification is a classical binary classification task to determine whether a given text is sexist or not, while Subtask 2 Sexism Categorisation is a finer-grained classification task focused on distinguishing different types of sexism.

To tackle this problem, we propose a novel approach XRCNN-EX by combining XLM-ROBERTA (XLM-R) [36] with a Text-based Convolutional Neural Network (TextCNN) [197] and infusing External lexical knowledge from HurtLex [73] to handle two subtasks of EXIST. Given the scarcity of semantic information in the commonly-used pooler output of XLM-R, XRCNN-EX aggregates the last 4 hidden states of XLM-R to obtain the representations with ampler semantic features. Then we construct a

TextCnn with 3 different kernels to capture various local features from Xlm-r, which decreases the memory cost with a smaller number of parameters and proceeds a faster training speed with lower computation compared to those Lstm-based models. Additionally, external knowledge from the domain-specific lexicon HurtLex is fed into the structure of Xrcnn in order to investigate the effectiveness of lexical information on performance.

In our experimental and official results, the basic architecture Xrcnn in our proposed model presents a notable achievement, while the performance of Xrcnn-Ex is comparatively unstable and inferior in the final submission. We discuss this case in Section 6.5. When it comes to the team ranking, we ranked 11^{th} in subtask 1 sexism identification and 4^{th} in subtask 2 sexism categorisation. In submission ranking, we ranked 14^{th} (accuracy score of 0.761) and 5^{th} (macro f1 score of 0.559) respectively.

6.2 EXIST: Task and Data Description

6.2.1 Task Description

The organisers of Exist proposed a shared task on automatic detection of multilingual sexist content on Twitter and Gab, including content in English (EN) and Spanish (ES). Two different subtasks were proposed:

- **Subtask 1 - Sexism Identification:** A binary classification task, where every system has to determine whether a given text (tweet or gab) is sexist or not sexist, where sexist content is defined as that which "is sexist itself, describes a sexist situation or criticises a sexist behaviour."
- **Subtask 2 - Sexism Categorisation:** Aiming to classify the sexist texts according to five categories of sexist behaviour including: "ideological and inequality", "stereotype and dominance", "objectification", "sexual violence" and "misogyny and non-sexual violence".

Predictions should be made on a mixed test set including content in both languages. Subtask 1 is evaluated in terms of accuracy, while Subtask 2 is evaluated using a macro-F1 score. Each participating team could submit a maximum of 3 runs.

6.2.2 Data Description

The EXIST dataset, provided by organisers, consists of 6,977 tweets for training and 3,386 tweets for testing, both of which include content in English and Spanish, and are manually labelled by crowdsourced annotators. In addition, the test set also includes 982 "gabs" from the uncensored social network Gab.com in order to measure the difference between social networks with and without "content control", Twitter and Gab.com respectively. Table 6.1 shows more details of the datasets provided.

Table 6.1: EXIST dataset description.

Subtask 1	**Training**		**Testing**		**Subtask 2**	**Training**		**Testing**	
	EN	ES	EN	ES		EN	ES	EN	ES
Sexist	1636	1741	1158	1123	**ideological-inequality**	386	480	333	288
					stereotyping-dominance	366	443	262	257
					sexual-violence	344	401	215	202
					misogyny-non-sexual-violence	284	244	198	202
					objectification	256	173	150	177
Non-sexist	1800	1800	1050	1037	**Non-Sexist**	1800	1800	1050	1037
Total						3436	3541	2208	2160
						6977		4368	
								Twitter: 3386 Gab: 982	

6.3 Methodology: XRCNN-Ex

In this section, we introduce our proposed model XRCNN-EX and experimental settings. Figure 6.1 shows the overall framework of the system we submitted to handle the two EXIST subtasks, which uses the pre-trained multilingual model XLM-R with the TextCNN and lexical features. We first obtain multilingual semantic information from the hidden state (the last 4 hidden layers) of XLM-R, and then concatenate

them together as the input to TextCnn for further feature extraction. External domain knowledge in the lexicon is incorporated into the basic structure of XRcnn and merged with the output of TextCnn. Finally, we pass the merged output features through a dense layer and utilise a softmax function for the final classification.

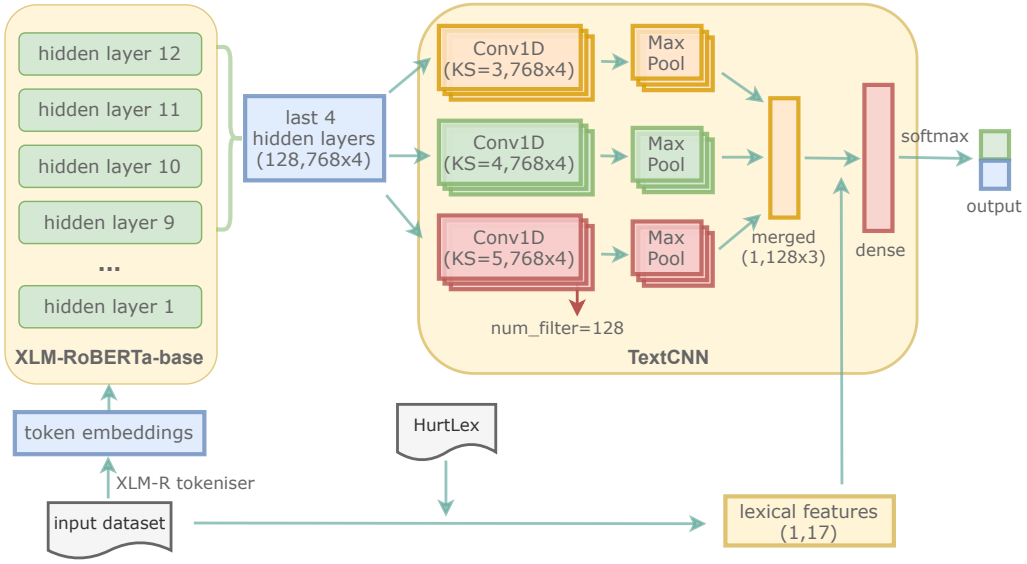

Figure 6.1: The overview of XRcnn-Ex architecture.

6.3.1 XLM-RoBERTa

Previous work with multilingual Masked Language Models (Mlm) has proved the effectiveness of pre-training large transformer models on multi-language corpora at once in the domain of cross-lingual understanding [35], such as multilingual Bert (mBert) [33] and cross-lingual language model (Xlm) [34]. These models have substantiated their superiority over supervised learning models in many Natural Language Processing (Nlp) tasks, especially in cases with limited training data. However, both mBert and Xlm are pre-trained on Wikipedia, leading to a relatively limited scale specifically for languages with poor resources. The Xlm-RoBerta model (Xlm-r) [36] has extended the way of pre-training MLM by scaling the amount of data by two orders

of magnitude (from Wikipedia to Common Crawl) and training on longer sequences (similar to RoBERTa [196]). It has been trained in more than 100 languages, leading to significant improvements on the performance of cross-lingual transfer tasks. In this work, we utilise XLM-R to address the multilingual EXIST dataset and extract semantic features of the whole text to deepen the understanding of the sentence and reduce the impact of noise.

The first token of the sequence in the last hidden layer of XLM-R is commonly used as the output for the classification task, while this output is usually not able to summarise abundant semantic information of the input sentence. Recent work by [215] indicates that richer semantic features can be learned by several hidden layers on top of BERT. In our system, we assume that some top hidden layers of XLM-R are also able to capture semantic information due to the similar architecture of XLM-R and BERT. Thus, we propose the model XRCNN as shown in Figure 7.1 for this task. Firstly, the input is processed by the XLM-R tokeniser and fed into the XLM-R model to get a list of hidden states. Then we gain deeper semantic features by integrating the last 4 hidden layers of XLM-R and feed it into TextCNN. The shape of the output is $n \times (d \times 4)$, where n is the length of the input sentence, and d is the dimension of each token in one hidden layer.

6.3.2 TextCNN

A Text-based Convolutional Neural Network (TextCNN) is a popular architecture for dealing with NLP tasks with a good feature extraction capability [197, 216]. The network structure of TextCNN is a variant of the simple CNN model. It is comparatively simpler than other neural networks and is able to reduce the number of dimensions of the input features, resulting in a smaller number of parameters, lower computational needs, and a faster training speed [216]. TextCNN utilises several sliding convolution filters to capture local textual features [197].

In our system, we use multiple 1D convolution kernels at a time for the convolution operation over the output of the last 4 hidden states from XLM-R. The output feature set is $X = [x_1, x_2, x_3, ..., x_n] \in \mathbb{R}^{n \times (d \times 4)}$. Let the window $x_{i:i+j-1} = [x_i, x_{i+1}, ..., x_{i+j-1}]$ refer to the concatenation of j words. A filter $w \in \mathbb{R}^{j \times (d \times 4)}$ is involved in the convolution process, applied to the window $x_{i:i+j-1}$ of j words to generate a new feature c_i:

$$c_i = f(w \cdot x_{i:i+j-1} + b) \tag{6.1}$$

where f is a non-linear function such as ReLU and $b \in \mathbb{R}^{(d \times 4)}$ is the bias. After the filter w slides across $[x_{1:j}, x_{2:j+1}, ..., x_{n-j+1:n}]$, a feature map is generated:

$$C = [c_1, c_2, ...c_{n-j+1}] \in \mathbb{R}^{(n-j+1)} \tag{6.2}$$

Then we apply the global max-pooling operation over the feature map C and take the maximum value $\hat{c} = max\{C\}$ to capture the most important feature for each feature map [217]. Features extracted by multiple filters are merged and fed into a dense layer.

6.3.3 Lexical Feature Induction

Currently, language models based on the transformer architecture have been popular among many NLP tasks in both monolingual and multilingual scenarios. But one of the drawbacks is that these models do not take any additional domain knowledge into consideration, like linguistic information from the domain-specific lexicon [218]. Bassignana et al. [73] introduce HurtLex, a multilingual lexicon containing offensive, aggressive, and hateful words and phrases in over 50 languages and spanning 17 categories [73]. The work by Koufakou et al. [37] incorporated lexical features based on the word categories derived from HurtLex to boost the performance of monolingual

BERT in such hate-related tasks, whereas there is no relevant study for the multilingual sexism scenario.

Given the scarcity of sexism-specific lexicons as well as the strong relation between those phenomena of offensive language and sexist language [17], we employ HurtLex for the induction of external lexical information to explore how the external lexical features affect the sexism detection performance. We extract 8,228 words for English and 5,006 for Spanish from HurtLex version 1.2, and construct multilingual lexical representations based on the HurtLex categories in both languages. There are 17 diverse categories, described with the number of terms in each language in Table 6.2. More specifically, we first generate a 17-dimensional lexical vector to count the frequency of each category. For instance, if a text includes 2 words in the category of derogatory words, the corresponding element in the lexical vector is supposed to be 2. Then we convert the lexical vector from the count frequency to Term Frequency–Inverse Document Frequency (TF-IDF) [198], indicating how significant a category is to a text in the corpus. Finally, we concatenate the TF-IDF lexical vector with the merged output of the TextCNN, and put it into the dense layer.

6.3.4 Output Layer

In order to prevent the model from over-fitting, we add the dropout after the dense layer, and then use a softmax function to obtain the label probability as the final output of the model.

6.3.5 Experimental Setting

Training Set Split

We use stratified sampling (StratifiedShuffleSplit) in the scikit-learn Python package for the cross-validation step instead of ordinary k-fold cross-validation to evaluate the model. Stratified Shuffle Split can create splits by preserving the same percentage for

Table 6.2: The category label, description and corresponding number of English and Spanish terms in HurtLex.

Label	Category Description	EN Terms	ES Terms
PS	negative stereotypes ethnic slurs	371	203
RCI	locations and demonyms	24	14
PA	professions and occupations	192	109
DDF	physical disabilities and diversity	63	36
DDP	cognitive disabilities and diversity	491	332
DMC	moral and behavioral defects	715	361
IS	words related to social and economic disadvantage	124	75
OR	plants	177	173
AN	animals	996	679
ASM	male genitalia	426	328
ASF	female genitalia	144	90
PR	words related to prostitution	276	165
OM	words related to homosexuality	361	213
QAS	with potential negative connotations	518	349
CDS	derogatory words	2204	1285
RE	felonies and words related to crime and immoral behavior	619	272
SVP	words related to the seven deadly sins of the Christian tradition	527	322

each target class as in the original training set. We set the number of splits to 5 and the ratio of the training set to the validation set to 9 to 1. For the EXIST training set, this led to a randomly sampled training set (6,279) and validation set (698). We present all performance scores in Section 6.4 based on the first split of training and validation sets.

Text Preprocessing

Since texts are obtained from Twitter and Gab, a pre-processing step is needed to maximise the features that can be extracted and to gain a unique and meaningful sequence of words, including removing non-alphabetic words, consecutive white spaces, and lowercasing all texts. As for special tokens in Twitter and Gab, we to-

kenise hashtags into separate words using the "wordsegment" Python package,[35] for example: *#HashtagContent* becomes *Hashtag Content*. URLs are replaced with the meta-token <URL> and user names are replaced with <USERNAME>. The text is subsequently tokenised using the corresponding XLM-R pre-trained tokeniser for both languages.

Model Parameter Setting

The parameters in each part of XRCNN-EX are shown below:

- XLM-R: we use XLM-ROBERTa-base pre-trained model, consisting of 12 hidden layers. We set the output hidden states in XLM-R config file to True in order to obtain different hidden states.
- TEXTCNN: we set the number of filters to 128 and three kernel sizes of 3, 4, and 5. RELU is the non-linear function used for convolution operation.
- Dense layer: we set the number of units to 768.

Training Process

During our training process, we use sparse categorical cross entropy as the loss function to save time in memory and computation. We use the Adam optimiser with a learning rate of $1e^{-5}$. We set the max sequence length to 128 and the dropout rate to 0.4. The model is trained in 7 epochs with a batch size of 32. All implementations are under the environment of Keras 2.5.0 and Tensorflow 2.5.0 with Python 3.7. Considering the unequal distribution of labels, we select the macro-averaged F1 score and accuracy score as our evaluation metrics for both subtasks.

[35] https://pypi.org/project/wordsegment/

6.4 Experiments and Results

In this section, we report our results in the two subtasks of the EXIST competition. We first conduct comparative experiments to delve into the optimal way of consolidating features from the hidden state of XLM-R, and then perform an ablation study of the whole architecture of XRCNN-EX to probe the contribution of its different components. All results are evaluated on the training and validation sets from the first split of original training data released by the EXIST. The official results in the EXIST shared task are presented and discussed finally.

6.4.1 Comparative Experiments for XLM-R Outputs

The pooler output is commonly utilised as the output of PLMs to address the classification task, which is generally lacking in sufficient and effective semantic information in the sentence representation [215]. More semantic features can be explored from different hidden states of models.

In our experiments, we consider both pooler output and hidden state as the outputs of XLM-R, as well as investigate the consequence of diverse aggregations of several hidden layers. These experiments are implemented on the basic model structure XRCNN and results are displayed in Table 6.3. It can be observed that integrating the last 4 hidden states of XLM-R yields better performance than other outputs on both subtasks, showing a notable increase in comparison with the pooler output. To be more precise, the model with only the pooler output performs better than the one combining the last 2 hidden layers in subtask 1 and the one with the last hidden layer in subtask 2. Nevertheless, it does not outperform the model absorbed in more than 2 hidden layers, which designates the constraint of the pooler output as the output features and the benefit of abundant semantic information in the hidden layer of XLM-R infused in our model.

Table 6.3: The XRCNN performance in different aggregations of hidden layers in XLM-R.

XLM-R Hidden Layers	Subtask 1		Subtask 2	
	Accuracy	Macro F1	Accuracy	Macro F1
Pooler Output	0.754	0.753	0.609	0.527
Last Hidden Layer	0.768	0.768	0.651	0.561
Last 2 Hidden Layers	0.749	0.747	0.645	0.565
Last 3 Hidden Layers	0.801	0.799	0.625	0.541
Last 4 Hidden Layers	**0.804**	**0.804**	**0.663**	**0.590**

6.4.2 Ablative Experiments and Results

Our proposed model XRCNN-EX combines the last 4 hidden states of XLM-R and the TextCNN with 3 kernels, then inducting extra lexical information. Several ablative experiments are implemented by removing certain components of XRCNN-EX to understand the contribution of each component. The following models are applied in this step:

- XLM-R Last 4 Hidden Layers: we aggregate the last 4 hidden states of XLM-R as the sentence representations of the input and put them into a simple linear classifier.
- FASTTEXT + TextCNN: we use the FASTTEXT embeddings trained on Common Crawl and Wikipedia in 157 languages [128] to convert the input data into word embeddings, and then feed them into a TextCNN.
- XRCNN: basic architecture of our proposed model.
- XRCNN-EX: our proposed model incorporating lexical embeddings.

Results of the ablation study are reported in Table 6.4. We can see that XRCNN and XRCNN-EX both achieve competitive performance, with noticeable improvements over the other two ablative models XLM-R Last 4 Hidden Layers and FASTTEXT+TextCNN. Moreover, XRCNN-EX achieves a slight improvement in subtask 1 but it does not

outperform XRCNN in subtask 2, which casts some doubt on the impact of extra lexical embeddings. We further discuss this in the Section 6.5.

Table 6.4: Ablation experiments for different components of XRCNN-EX.

Model	Subtask 1		Subtask 2	
	Accuracy	Macro F1	Accuracy	Macro F1
XLM-R Last 4 Hidden Layers	0.788	0.788	0.639	0.539
FastText+TextCNN	0. 751	0.750	0.622	0.528
XRCNN	0.804	0.804	**0.663**	**0.590**
XRCNN-Ex	**0.806**	**0.805**	0.657	0.543

6.4.3 Official Results in the EXIST Shared Task

Table 6.5 presents the official results of different runs we submitted to handle the two subtasks as well as the best scores for the EXIST shared task. For these two subtasks, we submitted the results of XRCNN and XRCNN-EX. The results of XRCNN led to better final scores than XRCNN-EX, obtaining better ranks 14^{th} in subtask 1 (accuracy score of 0.761) and 5^{th} in subtask 2 (macro f1 score of 0.559). For the team ranking, we ranked 11^{th} in subtask 1 and 4^{th} in subtask 2.

Table 6.5: Official results on the test set.

Model	Subtask 1				Subtask 2			
	Accuracy	Macro F1	Rank (runs)	Rank (team)	Accuracy	Macro F1	Rank (runs)	Rank (team)
XRCNN	0.761	0.761	14	11	0.643	0.559	5	4
XRCNN-Ex	0.756	0.756	18	12	0.635	0.546	13	10
Best score	0.780	0.780	-	-	0.659	0.579	-	-

6.5 Discussion

Our results show that the inclusion of the hidden state of XLM-R and TextCNN effectively improves the model quality of identifying sexist content, which is the most significant contribution of this work. However, results on the test set for XRCNN

model with lexical features demonstrate that the choice of lexicon words needs to be done more carefully, as they can lead to harming performance as is the case of XRCNN-Ex in the final scores. We foresee the need to further investigate the following variations to assess their impact on the performance:

- **Dataset variety:** The variety within the dataset impacts the effectiveness of semantic features. Lexical terms might be unevenly distributed across the training and test sets, leading to variations in the model's performance. For instance, sexist phrases that are contextually rich in the training set may not be adequately represented in the test set, affecting detection accuracy.

- **Term inconsistency between dataset and lexicon:** Terms in the dataset and the lexicon could be inconsistent. The hate-specific lexicon might not be capable of covering all hate-related terms encountered across different datasets. This gap suggests that while semantics can improve detection, its effectiveness is dependent on the comprehensiveness of the lexicon.

- **Linguistic characteristics:** Not all posts containing hateful terms are sexist necessarily, due to cases of polysemy or negation. Adding semantic knowledge into the model may exacerbate this issue.

- **Humour, irony and sarcasm:** Sexist posts with humour, irony and sarcasm are implicit and difficult to be identified. These posts often lack explicit hate-related terms, so adding external lexical knowledge is unlikely to be effective in addressing this issue.

- **Spelling variation:** Spelling variation is prevalent in social media [219]. Sensitive words sometimes use spelling variations to obfuscate and avoid detection, which do not match those normative words in the lexicon. Our XRCNN-Ex's reliance on a fixed lexicon sometimes fails to recognise these variations, suggest-

ing that incorporating advanced spelling variation detection techniques could enhance performance.

- **Quality of lexical features:** TF-IDF frequency features captured from the category of lexical terms might be comparatively sparse and lose information for specific terms. Lexical embeddings derived from pre-trained word embedding models could be beneficial as high-quality word embeddings can be learned efficiently thanks to low space and time complexity [220].

- **Approaches for lexicon induction:** Since the approach for lexicon induction might not fully absorb lexical information by simple concatenation between textual hidden features and lexical features, other forms of fusion can be tested, such as matrix multiplication [221] and cosine similarity [222].

Overall, our findings underscore the significant role of semantics in improving the detection of hateful content, particularly for context-dependent and implicit expressions of sexism. However, the effectiveness of these semantic features is influenced by various factors, including dataset variety, lexical consistency, and linguistic characteristics. Future work should explore advanced techniques for handling humor, irony, and spelling variations, as well as more sophisticated methods for integrating lexical information to further enhance model performance.

6.6 Conclusion

In this chapter, we propose a novel system called XRCNN-EX for multilingual sexism identification in English and Spanish social media. Our basic architecture XRCNN in XRCNN-EX, instead of only using the pooler output as the XLM-R's output to deal with the classification task, incorporates the last 4 hidden layers of XLM-R to gain deeper and richer semantic representations, which is fed into a faster classifier TextCNN.

Results in both validation and test sets indicate the effectiveness of using multiple hidden states with enriched semantic information and the capability of the TextCnn classifier on top of Xlm-r. In addition, we delve into the impact of integrating hate-related lexical embeddings into the system Xrcnn-Ex.

7

Retrofitting Sexism Domain-Aware Word Embeddings for Low-Resource Languages

In this chapter, we specialise the existing word embeddings with domain knowledge for one of the low-resource languages – Chinese. We develop a cross-lingual domain-aware semantic specialisation system to make the most of existing data to construct Sexist Word Embeddings (SexWEs), facilitating the performance of the sexism detection task for low-resource languages.

The chapter is organised as follows. In Section 7.2, we introduce how we investigate semantic specialisation for cross-lingual sexism detection and build sexism-specific word embeddings for Chinese. Then we describe experimental settings for the intrinsic evaluation of word similarity and the extrinsic evaluation of sexism detection in Section 7.3, and analyse results for both evaluations in Section 7.4. Section 7.5 provides further discussions based on SexWEs and experiments.

7.1 Introduction

The goal of sexism detection is to mitigate negative online content targeting certain gender groups of people. However, the limited availability of labelled sexism-related datasets makes it problematic to identify online sexism for low-resource languages. Rather than collecting new sexism data or building cross-lingual transfer learning models, we develop a cross-lingual domain-aware semantic specialisation system in order to make the most of existing data. Semantic specialisation is a technique for retrofitting pre-trained distributional word vectors by integrating external linguistic knowledge (such as lexico-semantic relations) into the specialised feature space.

To do this, we first structure linguistic constraints from external sexism-related semantic knowledge (e.g., BabelNet [223]) into different forms, including source constraints (English), target constraints (Chinese) and cross-lingual constraints. Then we project all source constraints into target constraints, and refine these projected target constraints by cleaning up the noise inside them. After that, various target constraint groups are incorporated together into the specialisation process to retrofit

pre-trained word embeddings to be domain-aware for the target language. Finally, we can monolingually employ our domain-specific SexWEs to the downstream task of social media sexism detection.

Furthermore, we verify the quality of our SexWEs in the intrinsic evaluation of word similarity, as well as the impact on sexism detection. Our results show that SexWEs achieves state-of-the-art (SOTA) performance on several word similarity benchmarks, outperforming all baseline classifiers on identifying sexism. Additionally, the visualisation of SexWEs with diverse constraints shows positive changes before and after the specialisation, and an ablation study also demonstrates the effectiveness of our proposed architecture for cross-lingual domain-aware specialisation. Our specialisation method enables us to specialise any type of distributional vectors in the target language with diverse constraints.

We publicly release our resources[36] to facilitate the integration of external lexical domain knowledge into distributional embedding models for other low-resource languages.

7.2 Methodology: SexWEs

We propose to build Sexist Word Embeddings (SexWEs) based on a cross-lingual domain-aware semantic specialisation system, inspired by the Cross-Lingual Specialisation transfer based on lexical Relation Induction (CLSRI) framework [97]. The objective is to incorporate awareness of the sexism domain into the semantic specialisation procedure to enrich domain-aware word embeddings (integrated sexism domain knowledge). We aim to specialise existing SOTA distributional word embeddings in a target language by utilising commonsense knowledge and multilingual domain knowledge from lexical constraints, where constraints are dominated by resource-rich source language and supplemented by a resource-poor target language. In our case, we opt

[36] https://github.com/aggiejiang/SexWEs

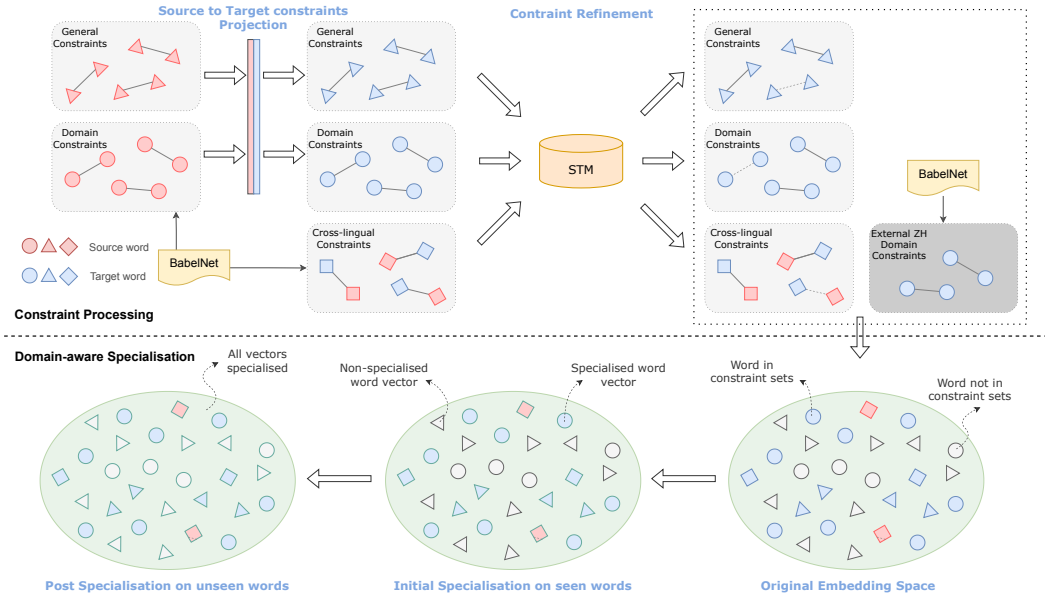

Figure 7.1: Overview of SexWEs. Constraint processing collects multilingual domain constraints, projects them across languages and filters noisy pairs. Domain-aware specialisation retrofits distributional word vectors in two steps: (1) utilise knowledge-aware constraints to specialise vectors on seen words; (2) learn and apply specialised mapping to the entire space.

for English (EN) as the source language L_{en} and Chinese (ZH) as the target language L_{zh}.

Our procedure can be split into two parts: constraint processing and domain-aware specialisation (see Figure 7.1). Firstly, constraint processing is to collect multilingual domain constraints, project source constraints across languages and clean up noisy constraints by transformation. Then we fuse the refined target constraints and external target constraints together as constraints C_{zh}^{group}, and execute monolingually initial specialisation and post-specialisation on existing distributional word vector space by employing well-handled constraints C_{zh}^{group} in the target language.

7.2.1 Constraint Processing

According to Mrkšić et al.'s [39] ATTRACT-REPEL methodology, linguistic constraints obtained from external sources are usually divided into two lexico-semantic groups:

- *ATTRACT constraints*: indicate word pairs with similar representations, e.g., synonyms (swearing and abuse, 咒骂 and 辱骂) or direct hypernym-hyponym pairs (woman and widow, 女人 and 寡妇);
- *REPEL constraints*: specify which word pairs should appear far apart in the vector space, e.g., antonyms (appreciation and disgust, 欣赏 and 厌恶).

Our constraints are grouped into five categories:

- English general constraints C_{en}^{g}
- English domain constraints C_{en}^{d}
- English general&domain constraints $C_{en}^{both} = C_{en}^{g} \cup C_{en}^{d}$
- Chinese domain constraints C_{zh}^{d}
- Cross-lingual EN-ZH domain constraints C_{cl}^{d}

English general constraints include words that are commonly and frequently used, while domain constraints refer to words related to the domain. In our case, we continue to use the existing general constraints [97] and extract domain constraints in both monolingual and cross-lingual scenarios. Except for C_{cl}^{d} constraints, the other four types of constraints all have ATTRACT and REPEL sets separately. This step focuses on processing source constraints and cross-lingual constraints while target constraints C_{zh}^{g}, C_{zh}^{d} or C_{zh}^{both} could be regarded as external constraints to facilitate specialisation performance in the next step.

In this constraint processing step, we first collect domain constraints into ATTRACT and REPEL sets from BabelNet[37] in source language C^d_{en} and target language C^d_{zh}, and project all constraints in source language C^d_{en} to those in target language $C^d_{zh'}$. Considering imperfect mapping and polysemy of C^d_{en} possibly leading to the incorrect meaning of $C^d_{zh'}$, these noisy constraints $C^d_{zh'}$ are filtered via a variant of Specialisation Tensor Model (STM) [224].

Monolingual and Cross-lingual Domain Constraints Collection

In order to extract monolingual domain constraints, we organise domain seed words from several domain-related lexical resources for both source C^d_{en} and target C^d_{zh} languages separately. Then we create domain constraint pairs via searching synonyms and antonyms in the same language for each seed word, and add a language tag before each word, such as (zh_ 歧视, zh_ 偏见).[38]

In addition to monolingual domain constraints, we also extract cross-lingual domain constraints C^d_{cl} based on domain seed words in the form of English-Chinese constraints. It will be taken into consideration such as explicit and implicit cross-lingual domain constraints. An explicit constraint refers to those English-Chinese constraints via direct translation and both words are explicitly domain-related (such as (en_prejudice, zh_ 歧视)), while an implicit constraint means two words cannot be directly translated from/to each other, because one word is domain-related in one language but another one could be domain-unrelated in another language if directly translated (such as (en_f*cking, zh_ 草)[39] and (zh_ 绿茶婊, en_angelic b*tch).[40]

[37] https://babelnet.org/

[38] 歧视 or 偏见 means an unfair and unreasonable opinion or feeling, especially when formed without enough thought or knowledge, such as prejudice or bias.

[39] The primary meaning of 草 is grass, only in certain occasions it may mean the same as f*cking.

[40] 绿茶婊 refers to girls who pretend to be pure and innocent but in fact are manipulative and scheming. It literally translates into green tea b*tch. The meaning of 绿茶婊 is similar to angelic b*tch.

All domain seed words are first directly translated[41] into explicit constraints, then we manually check and correct incorrectly translated word pairs to generate implicit constraints.

Source to Target Constraints Projection

Learning cross-lingual word embeddings (CLWEs) via supervised approaches shows good performance on the task of Bilingual Lexicon Induction (BLI) especially on typologically-distant language pairs like EN-ZH [225]. Recent work [225] has shown that Relaxed Cross-domain Similarity Local Scaling (RCSLS) [226], as a supervised system, achieves remarkable performance among competing models on the BLI task, and it has been applied to the word translation task in order to enhance the performance. So we leverage the RCSLS model to learn a linear cross-lingual projection matrix $\mathbf{W}_{en\_zh}$ between source and target word embeddings.

Given a set of source constraints C_{en}, each constraint is presented as a word pair (w^a_{en}, w^b_{en}). Since phrases exist widely in domain constraints C^d_{en}, phrase-level projection is also employed by averaging all word embeddings per phrase. We translate each word or phrase w_{en} in source constraints by looking for the nearest neighbour of its (averaged) embedding $\mathbf{x}_s$ in the projected target space. We project source constraints C_{en} into target constraints $C^d_{zh'}$ by using the projection matrix $\mathbf{W}_{en\_zh}$ to project source and target embeddings into a shared bilingual space $\mathbf{X}_{en\_zh}$.

Target Constraint Refinement

The shared bilingual space obtained by the cross-lingual projection matrix is far from perfect due to incorrect translation via the cross-lingual shared space and incorrect senses of polysemous words in L_{en}. Hence, noisy constraints could be generated via

[41] We use Google Translate https://translate.google.co.uk/.

projection-based approaches from source constraints C_{en} to target constraints $C^{d}_{zh'}$ [96].

Similar to the CLSRI framework, we aim to purify noisy constraints in $C^{d}_{zh'}$ by leveraging the STM to discriminate lexico-semantic relations within word pairs [224]. STM is a simple and effective feed-forward neural architecture that predicts lexical relations between word pairs by specialising input distributional word embeddings in multiple different projections and computing latent scores from these specialisation tensors for the final relation classifier. STM performs better, particularly for synonyms and antonyms, and also presents stable performance across languages [224]. We alter the multi-label STM classifier to a binary classifier,[42] and train five types of instances for STM:

- G_a-STM & D_a-STM: it predicts whether a word pair from general or domain constraints represents a valid ATTRACT constraint;
- G_r-STM & D_r-STM: it predicts whether a word pair from general or domain constraints represents a valid REPEL constraint;
- D_{cl}-STM: it predicts whether a pair of cross-lingual domain words represents a valid ATTRACT constraint;

7.2.2 Domain-Aware Specialisation

The step of domain-aware specialisation consists of monolingually retrofitting distributional word embeddings space in the target language L_{zh} by leveraging a group of target constraints, such as projected target constraints (e.g., $C^{g}_{zh'}$, $C^{d}_{zh'}$ or $C^{both}_{zh'}$) plus

[42] See more STM technical details in Glavaš and Vulić [224].

external target constraints (e.g., C^{g}_{zh}, C^{d}_{zh} or C^{both}_{zh}). The whole semantic specialisation process is similar to the CLSRI system [97]. Following the SOTA specialisation model ATTRACT-REPEL (AR) [39], we initially specialise the target distributional space to be domain-aware but limited to existing C_{zh} constraints. Then, based on the AR specialisation, we apply the SOTA post-specialisation model RetroGAN [95] to the entire vocabulary V_{zh}, including all the words seen and unseen in the target space. The following is a detailed description of the system and a brief outline of AR and RetroGAN models.

Initial Domain-Aware Specialisation

The group of target constraints C^{group}_{zh} to be specialised is a combination of projected target constraints $C_{zh'}$ from source constraints C_{en} and external target constraints C_{zh} from scratch, where $C^{group}_{zh} = C^{type}_{zh} \cup C^{type}_{zh'}$ and $type = \{g, d, both\}$. After the combination, C^{group}_{zh} includes two constraint subsets: ATTRACT constraints A_{zh} and REPEL constraints R_{zh}. The distance of each word pair (w^{a}_{zh}, w^{b}_{zh}) from A_{zh} and R_{zh} is refined between their corresponding embeddings $(\mathbf{x}^{a}_{zh}, \mathbf{x}^{b}_{zh})$ in the target distributional space.

The specialisation process is carried out via mini-batches of C^{group}_{zh}. Let $\mathcal{B}_A$ be a batch of vector pairs from A_{zh} and $\mathcal{B}_R$ the batch from R_{zh}. We define $\mathcal{T}_A(\mathcal{B}_A)$ and $\mathcal{T}_R(\mathcal{B}_R)$ as corresponding negative pairs for each $\mathcal{B}_A$ and $\mathcal{B}_R$. For each A_{zh} (or R_{zh}) constraint $(\mathbf{x}^{a}_{zh}, \mathbf{x}^{b}_{zh})$, we retrieve its closest (or farthest) vector pair as the negative constraint $(\mathbf{t}^{a}_{zh}, \mathbf{t}^{b}_{zh})$. Half of the negative constraints are selected based on their cosine similarity, and the other half are random negative samples.

The objective of AR retrofitting is to minimise the max-margin loss between target constraints and their corresponding negative samples, which includes three types of losses:

$$\mathcal{L}_{AR} = Att(\mathcal{B}_A, \mathcal{T}_A) + Rep(\mathcal{B}_R, \mathcal{T}_R) + Pre(\mathcal{B}_A, \mathcal{B}_R) \tag{7.1}$$

Specifically, $Att(\mathcal{B}_A, \mathcal{T}_A)$ enables target constraints in $\mathcal{B}_A$ closer together than those in the corresponding $\mathcal{T}_A$ by a ATTRACT margin δ_A:

$$\begin{aligned} Att(\mathcal{B}_A, \mathcal{T}_A) = \sum_{i=1}^{|\mathcal{B}_A|} [\mathcal{T}(\delta_A + \mathbf{x}^a_{zh_i}\mathbf{t}^a_{zh_i} - \mathbf{x}^a_{zh_i}\mathbf{x}^b_{zh_i}) \\ + \mathcal{T}(\delta_A + \mathbf{x}^b_{zh_i}\mathbf{t}^b_{zh_i} - \mathbf{x}^a_{zh_i}\mathbf{x}^b_{zh_i})] \end{aligned} \tag{7.2}$$

where $\mathcal{T}(x) = max(0, x)$ is the hinge loss function, and δ_A determines how much closer target constraints from A_{zh} are to each other than the distance to their corresponding negative examples. Analogously, $Rep(\mathcal{B}_R, \mathcal{T}_R)$ imposes constraints in $\mathcal{B}_R$ farther than their corresponding constraints in $\mathcal{T}_R$ based on a REPEL margin δ_R. Besides, $Pre(\mathcal{B}_A, \mathcal{B}_R)$ is the regularisation term to preserve the high-quality semantic information from $\mathbf{X}_{zh}$ by minimising the Euclidean distance between plain and specialised embeddings.

After AR specialisation, AR specialised space $\mathbf{X}^{'}_{zh} \in \mathcal{R}^d$ is generated from the initial distributional space $\mathbf{X}_{zh} \in \mathcal{R}^d$.

Cyclic Adversarial Post-Specialisation

The AR specialisation only works on the target words V^{seen}_{zh} that actually exist in C^{group}_{zh}, which indicates that the performance of initial specialisation can be semantically improved in terms of the overlapping vocabulary between explicit V^{seen}_{zh} and the vocabulary V_{zh} of the initial distributional space $\mathbf{X}_{zh}$. Post-specialisation learns the mapping from the initial specialisation space and propagates it to the rest of the vocabulary V^{unseen}_{zh} [92, 95].

RetroGAN enriches the existing adversarial post-specialisation model [94] to a CycleGAN-like architecture with a pair of Generative Adversarial Networks (GANs) [227]. The goal of RetroGAN is to learn a global specialisation mapping by balancing a combination of losses in both post-specialisation and inversion to ensure a unique one-to-

one mapping between the plain vector space $\mathbf{X}_{zh}$ and specialised AR space $\mathbf{X}^{'}_{zh}$ as conditioned by embeddings of seen words V^{seen}_{zh} from C^{group}_{zh} constraints. Then it propagates this global mapping to the entire distributional space of our target language $\mathbf{X}_{zh}$.

The model combines both cyclic and non-cyclic optimisation objectives, where the contrastive margin-based ranking loss with random confounders L_{MM} [94, 97] is used for both the generators and additionally for the cycle of generators:[43]

$$
\begin{aligned}
\mathcal{L}_{MM} = & \sum_{i=1}^{||V^{seen}_{zh}||} \sum_{j=1|j\neq i}^{k} \mathcal{T} \\
& [(\delta_{MM} - \cos(G(\mathbf{x}_{zh_i}), \mathbf{x}^{'}_{zh_i}) + \cos(G(\mathbf{x}_{zh_i}), \mathbf{x}^{'}_{zh_j}) + \\
& (\delta_{MM} - \cos(F(\mathbf{x}_{zh_i}), \mathbf{x}^{'}_{zh_i}) + \cos(F(\mathbf{x}_{zh_i}), \mathbf{x}^{'}_{zh_j}) + \\
& (\delta_{MM} - \cos(G(F(\mathbf{x}_{zh_i})), \mathbf{x}^{'}_{zh_i}) + \cos(G(F(\mathbf{x}_{zh_i})), \mathbf{x}^{'}_{zh_j}) + \\
& (\delta_{MM} - \cos(F(G(\mathbf{x}_{zh_i})), \mathbf{x}^{'}_{zh_i}) + \cos(F(G(\mathbf{x}_{zh_i}), \mathbf{x}^{'}_{zh_j}))]
\end{aligned}
\tag{7.3}
$$

where $G : \mathbf{X}_{zh} \rightarrow \mathbf{X}^{'}_{zh}$ is the generator that maps the plain vector space $\mathbf{X}_{zh}$ to the specialised space $\mathbf{X}^{'}_{zh}$, and $F : \mathbf{X}^{'}_{zh} \rightarrow \mathbf{X}_{zh}$ is the generator that does the opposite. L_{MM} makes a word vector generated from $\mathbf{X}_{zh}$ by generators closer to its gold-standard vector (e.g., specialised AR vector $\mathbf{x}^{'}_{zh} \in \mathbf{X}^{'}$) and different from any of k random confounders by a margin δ_{MM}, and then forces this constraint across the cycle.

[43] See more technical details of the RetroGAN and its losses in Colon-Hernandez et al. [95].

7.3 Experimental Setup

7.3.1 Initial Distributional Word Embeddings

As a starting point to build domain-aware specialised embeddings, we employ publicly available FASTTEXT word vectors [128] for both English and Chinese.[44] They provide 300-dimensional word vectors trained on Common Crawl and Wikipedia in 157 languages, using Continuous Bag-Of-Words (CBOW) with position weights. We execute the projection from source to target vector space via the supervised RCSLS method, searching 10 nearest neighbours in 10 iterations.

7.3.2 External Sexism Lexical Knowledge

To generate domain-specific constraints, we intend to use some lexical resources related to sexist domains to organise sexist seed words. However, due to the lack of external resources specifically addressing sexism, we select words from abusive language-related resources, where abuse is a superdomain of sexism [14].

For the source language (EN), we use (i) the hate speech lexicon *HurtLex*, containing 6,287 seed offensive, aggressive, and hateful words and phrases in over 50 languages [73], and (ii) the abuse lexicon by Wiegand et al. [218], which includes 2,989 words. For the target language (ZH), we use SexHateLex [20], a large Chinese sexism lexicon including 3,016 profane and sexually abusive and slang words and phrases.

7.3.3 Linguistic Constraints

Linguistic Constraints are present in the form of word/phrase pairs in the source language (EN) and the target language (ZH) for semantic specialisation, which is divided into three categories: source general constraints, bilingual domain (sexism)

[44] Other multilingual embedding models, such as LASER, Multilingual BERT and XLM-R, could be tested, however they are generally better suited for sentence-level embeddings.

constraints and cross-lingual constraints. We also combine general and domain constraints (in the same language) as another group of constraints. The number of constraints is summarised in Table 7.1.

- *Source General Constraints:* We follow the same English general constraints as used in previous work for the specialisation process [94, 97]. These general constraints involve the lexico-semantic relations from WordNet [228], Paraphrase Database (PPDB) [229] and BabelNet [223], which covers 16.7% of the 200K most frequent English words in the vocabulary of FASTTEXT embeddings.

- *Bilingual Domain Constraints:* To produce domain constraints, we employ the multilingual semantic network BabelNet on sexism-related seed words or phrases to extract synonyms and antonyms according to word sense tags. These constraints cover only 14.4% and 4.2% of the English and Chinese vocabulary from FASTTEXT.

- *Cross-lingual Domain Constraints:* Cross-lingual sexism-related (domain) constraints are English-Chinese pairs extracted via multilingual BabelNet based on domain seed words or phrases (e.g., en_hate, zh_ 憎恶).

		General	**Sexism**	**Both**
English	ATTRACT	640,435	130,445	768,294
	REPEL	11,939	501	12,148
Chinese	ATTRACT	-	6,353	-
	REPEL	-	32	-
EN-ZH	ATTRACT	-	189	-

Table 7.1: Collection of ATTRACT and REPEL constraints for source (EN) and target (ZH). Both are the aggregate and deduplicated set of general and sexism-related constraints.

7.3.4 Specialisation Approaches in Comparison

We compare our SexWEs specialisation on different types of constraints with three other semantic specialisation methods, implemented using the same FASTTEXT embeddings and both constraints used for our model SexWEs:

- ATTRACT-REPEL (AR): A SOTA retrofitting approach [39] to refine a distributional vector space by using ATTRACT (synonymy) and REPEL (antonymy) constraints.
- RetroGAN: A post-specialisation approach [95] by learning the mapping of AR and then extending an adversarial post-specialisation model based on the Auxiliary-loss Generative Adversarial Network (AuxGAN) [94] into a CycleGAN-like architecture [230] on the entire dataset.
- CLSRI: A specialisation Transfer via Lexical Relation Induction [97] transfers specialisation mapping from a resource-rich source language (English) to virtually any target language based on AR and AuxGAN with noisy constraints cleanup.

7.3.5 Hyperparameters in the Training Process

Constraints Refinement: STM

The STM model is adopted to predict lexical relations between constraints with 5 specialisation tensors, 300 neurons of the hidden layer and a 0.5 dropout value based on prior work [97]. During training, we set the batch size to 32 and the maximum number of iterations to 10, using Adam optimiser [231] with a learning rate of 0.0001.

Initial specialisation: AR

We preserve the hyperparameter settings for AR as used by Mrkšić et al. [39]. The margins for ATTRACT, REPEL and regularisation are 0.6, 0.0 and $1e^{-9}$, respectively. The Adagrad optimiser [232] is used with a 0.05 learning rate, the batch size is 50, and the maximum number of iterations is 5. The same configuration as the baseline AR.

Post-Specialisation: RetroGAN

We use two hidden layers with 2,048 units for the generator and the discriminator in each GAN of RetroGAN, adopting 0.2 and 0.3 dropout rates separately. We set the margin δ_{MM} to 1.0 and the number of negative samples to 25, utilising Adam optimiser with 0.1 learning rate. The number of training epochs is set to 10 and batch size 32, same as the baseline RetroGAN model.

7.4 Results and Analysis

We evaluate our SexWEs via both intrinsic evaluation of word similarity and extrinsic evaluation of sexism detection.

7.4.1 Intrinsic Evaluation: Word Similarity

The first experiment is to assess the quality of our specialised space of SexWEs via the word similarity task, which aims to evaluate the ability of the model to capture the semantic proximity and relatedness between two words.

Chinese Embeddings in Comparison

We adopt original FASTTEXT word vectors and retrofitted vectors by other specialisation approaches in comparison with our specialised embeddings infusing diverse

constraints.

Evaluation Setup

We employ three word similarity benchmarks, namely SimLex-999 (SL999) [233], Word-Sim-296 (WS296) [234] and WordSim-240 (WS240) [235]. WS296 and WS240 are Chinese datasets, while SL999 is an English dataset then translated into traditional Chinese by Su and Lee [236]. We convert it from traditional to simplified Chinese with chinese-converter.[45] The word pair coverage in the datasets is 975 of 999 for SL999, 230 of 240 for WS240, and 286 of 297 for WS296. The Spearman's rank correlation ρ is measured as the intrinsic evaluation metric, as it effectively captures the monotonic relationship between the ranked similarity scores, even if the relationship between them is not strictly linear [237]. Here we evaluate the relationship between the gold word pair similarity scores by annotators and the cosine similarity scores of the corresponding word embeddings from various vector spaces.

Results and Analysis

The results of word similarity tests are summarised in Table 7.2. Regardless of whether we plus external Chinese domain constraints or not, our specialised SexWEs basically outperforms the initial distributional vectors (0.039) and other cross-lingual specialisation models (0.027), indicating the effectiveness of incorporating domain constraints in source language during the cross-lingual transfer. And to the best of our knowledge, our results also surpass the Chinese word embeddings VCWE [238] that achieves the SOTA performances on WS240 and WS296.[46] By fusing external domain-specific target pairs, it also achieves better results for vector space specialisation. Moreover,

[45] https://pypi.org/project/chinese-converter/

[46] The VCWE results are 0.578 for WS240 and 0.613 for WS296, and it exceeds many competitive Chinese embeddings [239]. For more results, see https://chinesenlp.xyz/docs/word_embedding.html.

even without the infusion of sexist-related knowledge, our approach still outperforms the similarly structured model CLSRI, while noticeably exceeding two separate models of AR and RetroGAN, respectively. Although our SexWEs achieves a satisfactory performance on SL999 among all models, it can still be noted that there is no big gap compared to scores on the other two benchmarks, probably due to the translation issue from English to Chinese version or the conversion issue between traditional and simplified Chinese.

	SL999	**WS240**	**WS296**
FASTTEXT	.347	.546	.620
AR	.402	.521	.586
RetroGAN	.380	.572	.615
CLSRI	.384	.558	.627
SexWes	**.406**	**.586**	.608
w/o external	.394	.581	.624
only general	.389	.561	.623
only domain	.388	.563	**.637**

Table 7.2: Results of word similarity evaluation based on Spearman's rank correlation score ρ (average of 5 runs).

7.4.2 Extrinsic Evaluation: Sexism Detection

We next implement extrinsic evaluation to adjust our specialised SexWEs to a downstream binary classification task – sexism detection – which assesses the effectiveness of word embeddings with domain information.

Dataset

We use the only sexism dataset in Chinese, Sina Weibo Sexism Review (SWSR) [20], with posts labelled for sexism from the Sina Weibo platform. SWSR annotations are constructed at different levels of granularity, and we use the binary labels: sexist and non-sexist. We split the entire dataset into training and test sets in the ratio of 4 to

1. We further randomly select 20% of the training set as the validation set for the model fine-tuning process, and finally utilise the whole training set to evaluate model capacity on the test set. More details are shown in Table 7.3.

	Train	Validation	Test
Sexist	2244	561	288
Non-Sexist	4214	1053	609
Total	6458	1614	897
STR (%)	34.7	34.8	32.1

Table 7.3: Distribution of train, validation and test sets, sexist text rate (STR) in the SWSR dataset.

Sexism Detection Models Tested

We leverage a simple Text-based Convolutional Neural Network (TextCNN) [197] as our primary classifier, which is a popular architecture for dealing with Natural Language Processing (NLP) tasks with a good feature extraction capability [216], leading to a smaller number of parameters, lower computational needs, and a faster training speed [216]. TextCNN is fed with different vectors used in the intrinsic evaluation, or changed to other SOTA models for comparison to demonstrate the impact of our specialised embeddings on detecting sexist text. For vectors, we use the original and specialised word embeddings evaluated in the intrinsic experiments in combination with static BERT embeddings extracted from Chinese BERT.[47]

As baseline models, we use BERT [33] and a SOTA Chinese pre-trained model MacBERT,[48] which adopts Masked Language Model (MLM) as correction in BERT. MacBERT performs better than normal Chinese BERT[49] and other variants in some

[47] We extract contextualised BERT embeddings from the initial embedding layer of Chinese BERT trained on SWSR training set, using Huggingface BERT model "hfl/chinese-bert-wwm-ext".

[48] https://huggingface.co/hfl/chinese-macbert-base

[49] https://huggingface.co/bert-base-chinese

classification tasks [240].

Evaluation Setup

We use the Adam optimiser (0.0001 learning rate) and a maximum sequence length of 100 for all baseline models. TextCnn contains 128 units in the hidden layer with the dropout value 0.4, and we use Huggingface models "bert-base-chinese" (Bert) and "hfl/chinese-macbert-base" (MacBert). We train the TextCnn-based models for 100 epochs and Bert-based models for 4 epochs, using the same batch size of 32. Given skewed label distribution, we report the accuracy and macro F1 scores as the evaluation metrics.

Results and Analysis

We report the results for sexism detection in Table 7.4. We see that the classifier with our SexWes achieves the highest F1 and accuracy scores, outperforming all baseline classifiers and classifiers with baseline retrofitted embeddings, and most of our models with different constraints display better results than baselines. The classifier with our SexWes also exhibits stable performance with relatively small fluctuations in scores. Comparing baseline embeddings, there are notable improvements (0.093-0.135) in our SexWes compared to those using Fasttext word embeddings and popular Chinese embeddings Vcwe, and better performance than Bert embeddings. Additionally, our model slightly outperforms Bert-related models Bert and MacBert, but both of them present smaller fluctuations due to high stability. RetroGan shows the best results among all baseline specialisation models and outperforms all non-specialised embeddings, but it is still below our SexWes. Moreover, we can draw some conclusions that are in line with the intrinsic evaluation. That is, leveraging sexism-related constraints and external constraints in the target language for the cross-lingual specialisation process improves the detection of online sexism, and only using general

constraints also shows the effectiveness in this task compared to other specialisation baselines.

Model		F1-sex	F1-not	Macro-F1	Accuracy
Baseline embeddings	+FT	.483 (±.015)	.723 (±.044)	.603 (±.028)	.641 (±.040)
	+VCWE	.355 (±.149)	.796 (±.010)	.645 (±.071)	.682 (±.008)
	+BERT_emb	.573 (±.059)	.835 (±.009)	.704 (±.027)	.765 (±.006)
	+AR	.490 (±.025)	.840 (±.017)	.668 (±.009)	.770 (±.011)
	+RetroGAN	.622 (±.010)	.811 (±.056)	.717 (±.027)	.753 (±.044)
	+CLSRI	.638 (±.005)	.775 (±.010)	.707 (±.006)	.723 (±.007)
Baseline models	BERT	.641 (±.006)	.782 (±.008)	.711 (±.006)	.729 (±.007)
	MacBERT	**.658 (±.013)**	.789 (±.015)	.724 (±.013)	.739 (±.014)
SexWEs	**SexWes**	.626 (±.035)	**.849 (±.008)**	**.738 (±.016)**	**.786 (±.008)**
	w/o external	.627 (±.041)	.840 (±.044)	**.738 (±.024)**	.761 (±.034)
	only general	.622 (±.061)	.842 (±.011)	.732 (±.030)	.779 (±.012)
	only domain	.646(±.011)	.817 (±.056)	.733 (±.032)	.764 (±.046)

Table 7.4: Results of sexism detection with standard deviations (average of 10 runs).

Impact of Class Imbalance

Sexism or abuse tends to be the minority class in most datasets. In the case of the SWSR dataset, 65.5% are non-sexist instances [20]. Our SexWEs F1 score for the sexist class is 0.626, which is still clearly below the F1-not score of 0.849. We can also clearly observe that the F1 scores between sexist and non-sexist classes differ greatly, with the average F1 score of the non-sexist class being about 0.227 higher than that of the sexist class. This shows a negative impact of class imbalance on the sexism detection task, and the potential challenges that sexist texts may bring to the detection (see more in the subsection Qualitative Analysis).

Resampling and data augmentation techniques could be considered to mitigate the imbalance in the future [137, 241].

Qualitative Analysis

In addition to quantitative evaluation, we also conduct a qualitative analysis of some cases to assess the potential of SexWEs for sexism detection as well as the challenges, and examples are presented in Table 7.5.

Text	TextCNN+FT	BERT	SexWEs	Ground Truth
1. 只要你为女性发声，只要你有独立意识，你就是“女拳”。 Translation: As long as you speak for women and have independent thoughts, you are labeled as “feminist”.	Non-Sexist	Non-Sexist	Sexist	**Sexist**
2. 我也纳闷。这踮明明是德普被前妻家暴，还能遭来这么些个女人洗地、好像她们自己这辈子都圣母得没暴力过男人。 Translation: I am also puzzled. This was clearly Depp being abused by his ex-wife, yet so many women defend him, as if they have never been violent to men in their lives.	Non-Sexist	Non-Sexist	Sexist	**Sexist**
3. 尊重驴不带套的权力，意外怀孕的权力，尊重就 vans。 Translation: Respect the rights of women without wearing condoms and unintended pregnancies, that’s it.	Non-Sexist	Non-Sexist	Sexist	**Sexist**
4. 学历高的估计更厉害，从道理上说服你，不然就身体上睡服你。 Translation: Males with higher education may be better at persuad -ing you or f*cking you.	Non-Sexist	Non-Sexist	Non-Sexist	**Sexist**
5. 田园女权，女拳师，极端女权，是我是我都是我。 Translation: Pastoral feminist, female boxer, extreme feminist, it is all me.	Sexist	Sexist	Sexist	**Non-Sexist**

Table 7.5: Examples with predictions by three models: TextCNN + FASTTEXT embeddings (TextCNN+FT), BERT, and TextCNN + specialised embeddings (SexWEs), along with ground truth labels.

To investigate how our SexWEs with semantic word pairs help in detecting complex forms of sexism, we focus on examples where the model with specialised embeddings correctly identified sexist content while the other models (BERT and TextCNN with FASTTEXT embeddings) did not.

In example (1), the term “女拳” (feminist), a homophone of “女权”, is used derogatorily to label women who advocate for themselves and exhibit independence, where “女拳” is negative but “女权” is neutral. This labeling of “女拳” is intended to demean and diminish the value of women’s voices. The semantic relationship between terms such as “女性” (women) – “女拳” helps to identify the negative connotation of sexism. By understanding these relationships, the model can recognise the derogatory intent behind labeling independent women as “feminist”, something that might be missed

without semantic context. Besides, the text in example (2) implies that women are generally violent and hypocritical in domestic violence situations, making sweeping assumptions about women's behaviour. Semantic pairs, such as "家暴" (domestic violence/abuse) – "暴力" (violent) and "家暴" – "女人" (women), are crucial for understanding the text's underlying assumptions and biases, highlighting how women are unfairly stereotyped based on isolated incidents. This demonstrates that recognising the semantic connections between these terms allows the model to detect the sexist more effective.

Furthermore, when looking at predicted examples from BERT and classifiers with original embeddings and our SexWEs, we see some recurrent types of misclassification as below.

(i) Implicit sexism: Humour, irony and sarcasm are difficult to be identified. Example (3) is sexist irony without an explicitly abusive expression. The model with SexWEs successfully deemed it sexist, while the others failed.

(ii) Informal and code-mixed expressions: Example (3) is a Chinese-English code-mixed text, and the slang word "vans" in English has a similar pronunciation as "完事了" (that's it) in Chinese. "驴" usually refers to "donkey", but it is commonly used in sexist expressions that offend women.[50]

(iii) Implicit attack target: The attack target might not explicitly appear like in example (4). All models failed to predict it as sexist text. It demeans the group of highly educated males, but the target can only be guessed from the context.

(iv) Homophones: Homophones are common in sexist speech to convey abusive connotations, or to obfuscate and avoid detection. "说服" and "睡服" have the same pronunciation in (4). "说服" is a general term that means persuade or convince, and "睡服" is a homophonic word with a similar meaning to persuade someone by f*cking.

[50] "驴" comes from "婚驴" (marriage donkey), and is intended to depict the image of "women who are as stupid as donkeys in marriage, deprived of a lot of benefits, but still enjoy silly happiness".

(v) Overuse of explicit sexist words: Sexist words might be overused in one text, leading to the over-dependence of the model on these words, while sexist targets in posts are confounding and hard to identify. All models failed in example (5), and we see that the model can easily deem a text sexist if it contains many sexist words, despite not having a specific targeted individual or group.

7.5 Discussion

7.5.1 Visualisation of Word Embeddings

Figure 7.2: t-SNE visualisations of SexWEs word embeddings. Each colour group indicates a Chinese domain word with its 20 neighbours generated from original FASTTEXT vectors. There is a total of 6 seed words selected, namely purple for 女人 (woman), blue for 性侵 (sexual assault), skyblue for 强奸 (rape), green for 下贱 (b*tchy), orange for 傻 (stupid), and red for 责骂 (scold). The averaged local distance of word clusters (local_dist) is measured based on the t-SNE space.

We visualise both original FASTTEXT embeddings and various specialised SexWEs embeddings. We select six sexism-related seed words and gather each seed word with its 20 nearest neighbors from the initial word vector space, to explore changes

in these domain word groups during our specialisation process. Figure 7.2 shows the visualisation of word embeddings with dimensional reduction by t-distributed Stochastic Neighbour Embedding (t-SNE) [242]. To further investigate the semantic shift between different word vector spaces [243], we measure the average cosine distance between a seed word and its neighbours in each local word cluster, and average distances among the six clusters to obtain the overall distance in the space. The local distance is presented in subplot titles of Figure 7.2.

Looking at both the spatial range of visualised word clusters and local distances, we can observe that all specialised groups of domain words become more independent and get closer from the original distributional vector space in Figure 7.2 (a) to any of our specialised vector space (see Figure 7.2 (b)-(f)), which illustrates the benefit of our specialisation method. After the specialisation process with English constraints, the distance of word clusters shows a significant decrease, further decreasing after adding external Chinese constraints. For word embeddings that incorporate more domain information (Figure 7.2 (e) and (f)), the connections between words in each cluster become stronger, compared to embedding spaces that are only retrofitted with knowledge of general constraints in Figure 7.2 (d). Furthermore, after adding external Chinese constraints, the vector space specialised only with domain knowledge becomes more contiguous (see Figure 7.2 (e) to (b)), while the spaces specialised by both constraints are relatively sparse (see Figure 7.2 (f) to (c)). This opposite change may be caused by perturbations of commonsense knowledge, since general constraints outnumber domain constraints.

7.5.2 Ablation Study

To evaluate different components, we perform a study of the following ablated models of SexWEs: *(i) Remove phrase-level projection:* Only project source to target constraints on word level; *(ii) Remove constraint refinement:* Directly project tar-

get constraints into the specialisation without refinement; *(iii) Remove* R*etro*GAN *post-specialisation:* A variant of SexWEs without the RetroGAN.

	Intrinsic			Extrinsic	
	SL999	**WS240**	**WS296**	**Macro-F1**	**Acc.**
SexWes	**.406**	**.586**	.608	**.738**	**.786**
w/o phrase	.404	.571	**.611**	.726	.778
w/o refinement	.390	.536	.591	.713	.768
w/o RetroGAN	.398	.529	.594	.704	.760

Table 7.6: Results for SexWEs and ablative methods.

In Table 7.6, we can see that our model outperforms all ablated models, which demonstrates the important contribution of all components. Although phrase-level constraint processing in the projection step does not significantly improve the quality of embeddings, this step validates the positive impact of doing domain-related phrase mapping on identifying sexism. The results also highlight the effectiveness of STM in refining the noisy lexico-semantic relations between constraints compared with the one without constraint refinement. Furthermore, we can validate the capability of RetroGAN post-specialisation step to efficiently apply the retrofitting mapping to full word vector space when compared to specialised word vectors without post-specialisation step.

7.5.3 Performance versus Complexity Trade-off Analysis

According to experimental results, the overall performance of SexWEs fine-tuned by our cross-lingual domain-aware specialisation system shows 0.004-0.065 correlation score improvement in word similarity benchmark and 0.014-0.135 F1 score improvement in sexism detection. The results of both intrinsic and extrinsic evaluations demonstrate the effectiveness of specialised word vectors compared to pre-trained word vector baselines, and show improved performance over all other specialisation systems with similar model complexity. Compared with BERT-related baselines, our

SexWEs is based on a simple TextCNN architecture and still achieves a slight increase in the performance of detecting sexist content, showing further potential for more advanced and robust networks. Furthermore, we only need to train once to construct sexist word embeddings. Instead of only using it for sexism detection, it can also be reused to study sexism-related issues. Only by collecting new constraints, the methodology of building the cross-lingual specialisation system can be further transferred to other low-resourced domains to detect abnormal behaviours online.

7.6 Conclusion

In this chapter, we propose an effective system for cross-lingual domain-aware semantic specialisation by injecting external constraints referring to sexist terms in both source and target languages. It can effectively tackle sexism detection for low-resource languages. We report notable performance of SexWEs in both intrinsic and extrinsic evaluations, visualising the positive trend of word embeddings during the specialisation, as well as through an ablation study. However, we only observe a modest improvement after adding cross-lingual constraints, potentially due to its limited size.

8

Conclusion

In this final chapter, we will recapitulate the proposed methods in Section 8.1, summarise our main contributions to the research field in Section 8.2, and provide an outlook into the future directions of the research in Section 8.3.

8.1 Synopsis

In this thesis, we have studied the problem of applying cross-lingual transfer learning (CLTL) techniques to automatically identify sexist hate speech across languages.

Chapter 3 provides a comprehensive overview of existing multilingual hate speech datasets, linguistic resources, and approaches utilised in CLTL. We also summarise and list existing challenges in the field regarding languages, datasets, and methods.

Chapter 4 broadens the scope of sexism detection by considering the Chinese language on Sina Weibo. We propose the first Chinese sexism dataset – Sina Weibo Sexism Review (SWSR) dataset – as well as a large Chinese lexicon SexHateLex made of abusive and gender-related terms. We introduce our data collection and annotation process, and provide an exploratory analysis of the dataset characteristics to validate its quality and to show how sexism is manifested in Chinese. The SWSR dataset provides labels at different levels of granularity including (i) sexism or non-sexism, (ii) sexism category and (iii) target type, which can be exploited, among others, for building computational methods to identify and investigate finer-grained gender-related abusive language. We conduct experiments for the three sexism classification tasks making use of state-of-the-art (SOTA) machine learning models, providing a benchmark for sexism detection in the Chinese language, as well as an error analysis highlighting open challenges needing more research in Chinese Natural Language Processing (NLP).

Chapter 5 investigates the cross-lingual hate speech detection task, and proposes a cross-lingual capsule network learning model coupled with extra domain-specific lexical semantics for sexism (CCNL-EX). It is a two-parallel framework, enriching

input information in both source and target languages. Our model achieves SOTA performance on benchmark datasets from AMI@Evalita2018 and AMI@Ibereval2018 involving three languages: English, Spanish and Italian, outperforming SOTA baselines on all six language pairs.

Chapter 6 proposes an architecture made of the last 4 hidden states of XLM-RoBERTa (XLM-R) and Text-based Convolutional Neural Network (TextCNN) with 3 kernels. Our model also exploits lexical features relying on the use of new and existing lexicons of abusive words, with a special focus on sexist slurs and abusive words targeting women. This work participated in the first shared task on sEXism Identification in Social neTworks (EXIST) at IberLEF 2021 [214], which provides a benchmark sexism dataset with Twitter and Gab posts in both English and Spanish, along with a task articulated in two subtasks consisting in sexism detection at different levels of granularity: Sexism Identification and Sexism Categorisation. Our model ranked 11^{th} and 4^{th} respectively in two subtasks among all the teams on the leaderboard, clearly outperforming the baselines offered by EXIST.

Chapter 7 addresses the task of automatic sexism detection in social media for one low-resource language – Chinese. Rather than collecting new sexism data or building CLTL models, we develop a cross-lingual domain-aware semantic specialisation system to make the most of existing data, by leveraging semantic resources for sexism from a high-resource language (English) to specialise pre-trained word vectors in the target language (Chinese) to inject domain knowledge. We demonstrate the benefit of our sexist word embeddings (SexWEs) specialised by our framework via intrinsic evaluation of word similarity and extrinsic evaluation of sexism detection. Compared with other specialisation approaches and Chinese baseline word vectors, our SexWEs shows an average score improvement of 0.033 and 0.064 in both intrinsic and extrinsic evaluations, respectively. The ablative results and visualisation of SexWEs also prove the effectiveness of our framework on retrofitting word vectors in low-resource

languages.

8.2 Summary of Contributions

The innovative aspect of this work lies in the application of transfer learning techniques to the task of hate speech detection in a cross-lingual setting. At the outset of our research, there was a very limited body of work in this area: only five studies focused on general hate speech [40–44], with just two of these addressing sexism [40, 42]. The lack of exploration in the field emphasises the novelty and importance of our work. The work comprised in this thesis contributes significantly to understanding and leveraging transfer learning for detecting cross-lingual sexist hate speech. It specifically highlights the following key contributions regarding our research objectives:

1. **Overcoming Resource Limitations in Target Languages:** One of the main challenges in cross-lingual transfer is the lack of adequate multilingual resources, especially for low-resource languages. To address the scarcity of resources in target languages such as Chinese, we design a data collection pipeline and annotation guidelines for creating a sexism dataset, including the formulation of what constitutes sexism in the context of an understudied language and culture, Chinese. Then we follow our pipeline and guidelines to construct and release the first Chinese sexism dataset (SWSR) to our knowledge (§4). The rich features of our SWSR dataset, including weibo contents, weibo reviews and basic user information, make it possible to detect sexist content with various approaches for better performance and interpretability, as well as enable a contextual analysis of sexism. Our dataset also provides a hierarchical taxonomy for the sexism category and the type of target of sexist comments, which enables finer-grained investigation of sexist texts. Besides, we integrate existing lexi-

cal resources and sexism-related terms to build a lexicon SexHateLex including 3,016 sexist and abusive terms. Our SWSR dataset and SexHateLex lexicon are publicly available,[51] achieving around 17.5k downloads so far. These resources will be supportive of Chinese sexism detection, and these development methods can be applied to other low-resource languages.

2. **Bridging Discrepancies Between Source and Target Languages:** The approach should overcome linguistic and cultural discrepancies between source (high-resource languages) and target (low-resource languages) settings. To do this, we utilise machine translation tools (e.g., Google Translate) to generate parallel datasets between source and target languages to enrich the input features across diverse languages for our CCNL-EX model (§5.3). Additionally, we leverage the Relaxed Cross-domain Similarity Local Scaling (RCSLS) model [226] to learn a linear cross-lingual projection between source and target word embeddings, strengthening connections between language pairs (§7.2).

3. **Transferring Domain Knowledge at Different Hierarchies of NLP Models:** The research will focus on developing various models to achieve domain knowledge transfer at different levels: instance, feature and model levels. For the instance level transfer, we apply mapping approaches between diverse languages to generate pseudo texts and labels (§5.3, §7.2). For the feature level transfer, we develop sexism embeddings in the target language via a cross-lingual specialisation technique, retrofitting word vectors in the source language with multilingual semantic relations (§7). This is the first study on semantic specialisation for cross-lingual abusive language detection, and our resources are publicly released.[52]. For the model level transfer, we achieve it by building the first sexism detection model that incorporates capsule networks in a

[51] http://doi.org/10.5281/zenodo.4773875
[52] https://github.com/aggiejiang/SexWEs

cross-lingual setting (§5), and proposing an XLM-R-based model with multi-layer features extracted to obtain richer semantic information for multilingual sexism detection (§6).

4. **Enhancing Model Understanding with External Knowledge:** To make models more domain-aware and language-aware, incorporating external resources can be of great benefit to improve model performance in detecting hate speech and sexism. However, at the start of our research, there were limited resources targeting hate speech and sexism, specifically in low-resource languages. Therefore, we build a new domain lexicon (SexHateLex) including 3,016 sexist and abusive terms by integrating existing lexical resources and sexism-related terms (§4.4). We also investigate the effectiveness of infusing hate-related lexicons and sexism-related semantic knowledge (e.g., BabelNet) into pre-trained word embeddings or pre-trained language models (§5.2.2, §6.3, §7.2). When selecting the external resource (such as lexicon, ontology, knowledge graph, word-pairs, etc.), we should consider its relevance to the target domain, coverage and completeness, quality, cultural sensitivity, and update frequency. A well-chosen lexicon can improve the task efficiency, while a poorly selected one may lead to missed detections or incorrect classifications, undermining effectiveness.

   Furthermore, our qualitative analysis (§5.4.3, §6.5, §7.4.2) demonstrates that infusing external knowledge can deepen the semantic understanding of language nuances, which is essential for identifying hateful and sexist content. This approach allows models to recognise derogatory terms and hate speech across various languages as well as enables models understand cultural contexts and implicit hate, thereby identifying terms that might otherwise be overlooked. Additionally, ensuring the consistency of lexicon words with the dataset and addressing spelling variations could further enhance the efficiency of utilising external knowledge.

5. **Evaluating and Refining Model Performance Across Languages:** To prove the effectiveness and robustness of our proposed methods, comparative experiments are conducted to evaluate our proposed cross-lingual models against diverse baseline models. We perform an exploratory analysis to validate the quality of our SWSR dataset and provide benchmark results among SOTA deep learning and pre-trained models (§4.6). We show the SOTA performance of CCNL-EX compared to ten baselines among different language pairs, and perform a comparative study looking into the impact of each layer on our model (§5.4). Then we conduct comparative experiments to delve into the optimal way of consolidating features from the hidden state of XLM-R, and then perform an ablation study of the whole architecture of XRCNN-EX to probe the contribution of its different components (§6.4). Besides, our specialised sexist embeddings (SexWEs) achieve SOTA performance on word similarity benchmarks compared to different specialisation transfer methods and the Chinese sexism detection task compared with all Chinese baseline embeddings (§7.4).

6. **Analysing Trends in Cross-Lingual Hate Speech Detection:** Given the growing trend of analysing hate speech across languages, we conduct the first systematic review of recent studies in the field of cross-lingual hate speech detection. This will involve surveying the existing 67 papers according to diverse aspects: multilingual datasets employed, cross-lingual resources leveraged, levels of transfer, and cross-lingual strategies applied. We also highlight current challenges in this field (§3). In addition, we present two comprehensive tables (§A) containing multilingual datasets and CLTL techniques used in surveyed papers respectively to facilitate easy comparison and discovery of related works.

8.3 Future Directions

The growing demand for advanced cross-lingual hate speech and sexism detection is evident, yet the complexities of language and culture continue to pose ongoing challenges. This section outlines promising research directions based on our work, aiming to enhance research capabilities in this critical area.

8.3.1 Dataset Creation

There is a pressing need to collect and annotate comprehensive and balanced datasets, especially in low-resource languages [150, 171, 181, 244]. While zero-shot cross-lingual transfer has its limitations, even a small amount of target data can substantially enhance model fine-tuning [99]. To avoid dataset bias, such datasets can span multiple languages, dialects, cultural contexts, and diverse topics of hate speech [82]. More diverse strategies for data collection, such as sampling from various platforms and user demographics, can yield richer and more annotation-worthy data [99]. Additionally, the creation of domain-specific evaluation datasets, such as XHATE-999 [84] and Multilingual HateCheck (MHC) [183], provides a multifaceted and precise perspective for assessing cross-lingual detection outcomes.

8.3.2 Data Annotation

Refining data annotation strategies will be instrumental. Instead of annotating all collected data at once, iterative annotation strategies can be used to start with smaller subsets, ensuring efficiency and encouraging diverse data collection [99]. In addition, semi-supervised learning methods can efficiently use data and annotator resources and mitigate data scarcity issues. For example, deliberate selection strategies, such as active learning, can enhance annotation performance with fewer target-language entries, mainly annotating data points that the model finds most ambiguous [99]. To

further address hate speech ambiguity and sensitivity, incorporating feedback from human experts can be invaluable [100]. It would also be better for annotators to follow ethical guidelines and receive appropriate training and support, avoiding potential harm to them.

8.3.3 Integration of Additional Features

Integrating additional features extracted from various resources has been identified as an efficient strategy to enhance cross-lingual model performance. **Typographic features**, including capitals, punctuations, and emojis, offer a language-agnostic and domain-agnostic perspective to hate speech detection. Emojis, associated with emotional expressions, have been highlighted for their ability to help identify online hate due to their shared meanings across languages [134, 162, 163]. **Stylometric features** are also robust indicators of hate speech, as the writing style of toxic content can be correlated to social media users [163]. In addition, **domain-specific features** are a critical aspect in additional features to assist the knowledge transfer process, especially when addressing linguistic nuances like metaphors, metonymy, and informal expressions. The multilingual lexicon, HurtLex, has been highlighted for its effectiveness in multiple studies [17, 29, 40, 120]. Since general-purpose multilingual embeddings may not effectively capture some hate-specific patterns, some researchers also utilise hateful word pairs to construct domain-aware or semantic-based representations across languages [89, 160, 161]. Furthermore, infusing **cultural features** underscores the significance of language-specific knowledge for geographically sensitive tasks like ours. The creation of cultural-aware models, informed by multidisciplinary studies involving anthropologists and sociologists, can provide a richer understanding of cultural dimensions in hateful content [31]. Such work can produce different forms of cultural features, such as switching pattern matrix [152], bilingual language pairs [161], cross-cultural similarities [89], and social dynamics among users [79].

8.3.4 Multilingual Pre-Trained Language Models

Cross-lingual hate speech detection has witnessed the rise of Multilingual Pre-trained Language Models (PLMs) as a leading technique, consistently achieving SOTA results. These multilingual PLMs can be applied to low-resource target languages without further training, thus bridging the data availability gap between high- and low-resource languages. However, there are significant computational costs during training and fine-tuning due to the large number of parameters [27]. Future research is geared towards optimising these multilingual PLMs for enhanced efficiency and interpretability. Additionally, two strategies are emerging to improve the generalisability and scalability of multilingual PLMs so that they can scale to handle large datasets or multiple languages simultaneously [74]. The first emphasises pre-training models using data from relevant sources (such as social platforms and hateful domains), as demonstrated by XLM-T [166] and AbuseXLMR [168]. The second strategy focuses on models based on specific low-resource languages or those from similar language families, as seen with MuRIL for Indic languages [146] and AraBERT for dialects [181].

8.3.5 Cross-lingual Training Strategies

Apart from the refinement of multilingual PLMs and the integration of multilingual knowledge graphs and semantic networks into the training process, many innovative cross-lingual training strategies are worth exploring. **Adaptive training techniques** can dynamically adjust to the specificities of different languages and dialects, as well as bridge the gap between hate speech detection task and PLMs. By leveraging meta-learning or few-shot learning, models can be trained to quickly adapt to new low-resource languages with a few labelled data [178]. Furthermore, collaborative **multi-task learning** can make models more generalisable in cultural variations and resistant to overfitting by training them and sharing representations on multiple auxiliary tasks simultaneously across different languages or similar NLP tasks

[166]. The use of **adversarial training**, where models are trained against adversarial examples, can also enhance their robustness and generalisation across languages. This is especially important for hate speech, as attackers often use subtle language tricks to obfuscate and bypass detection [30, 148, 149]. Moreover, **cascade learning** can greatly bolster model performance and enhance out-of-domain generalisability by initially training models on source data or more diverse datasets and subsequently fine-tuning the results by simultaneously or progressively incorporating target language data [99, 180]. Besides, Vitiugin et al. [100] highlight that **human feedback** can be of great value to help detect hate subtleties and phrases during model training, Ranasinghe and Zampieri [173] explore **language-specific preprocessing** like segmentation in morphologically rich languages such as Arabic and Turkish, and **multi-source training**, with augmented datasets, is also able to capture the diversities and commonalities of linguistic patterns across languages [42, 163, 177].

8.3.6 Application of Large Language Model

The emergence of advanced Large Language Models (LLM), such as GPT-3 [245], GPT-4 [246], Fine-tuned LAnguage Net (FLAN) [247] and Large Language Model Meta AI (LLaMA) [248], has opened up new horizons for cross-lingual hate speech detection, thanks to their profound linguistic understanding and emergent abilities [249]. Instruction fine-tuned LLM with prompting have shown promise in detecting hate speech across both mono and multilingual contents without the need for language-specific fine-tuning [26]. The efficacy of LLM can be further enhanced through well-designed, task-specific prompts [250–252]. By incorporating informative task descriptions and input-label demonstrations, these prompts can guide LLM to achieve superior detection outcomes. Furthermore, the inherent multi-tasking capabilities of LLM allow them to excel not only in hate speech detection but also in some related NLP tasks [253]. To address the challenge of data scarcity, LLM offer a potential solution by gen-

erating synthetic hateful texts, especially for languages with limited resources [254]. Beyond detection, LLM can also produce explanations for identified hate speech, enhancing both model performance and its interpretability [255]. While the potential of LLM in this domain is evident, current research is still in its infancy. Only a handful of studies have delved into the impact of LLM on hate speech detection, and only one focuses on multilingual scenarios. As such, there are still broad prospects for future research work to be explored.

References

[1] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. 51(4), jul 2018. ISSN 0360-0300. URL https://doi.org/10.1145/3232676.

[2] Anusha Chhabra and Dinesh Kumar Vishwakarma. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28, 2023.

[3] Rights for Peace. What is hate speech? https://www.rightsforpeace.org/hate-speech. Accessed: 2023-10-24.

[4] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

[5] Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117, 07 2019. ISSN 0007-0955. URL https://doi.org/10.1093/bjc/azz049.

[6] UNSPAHS. UN strategy and plan of action on hate speech. https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml, 2019. Accessed: 2023-10-24.

[7] Richard Ashby Wilson and Molly K Land. Hate speech on social media: Content moderation in context. *Conn. L. Rev.*, 52:1029, 2020.

[8] Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590, 2023. doi: 10.1109/ACCESS.2023.3258973.

[9] Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A Hale, and Paul Röttger. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. *arXiv preprint arXiv:2404.17874*, 2024.

[10] Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.236.

[11] Colin Dwyer and Andrew Limbong. 'go back where you came from': The long rhetorical roots of trump's racist tweets. https://www.npr.org/2019/07/15/741827580/go-back-where-you-came-from-the-long-rhetorical-roots-of-trump-s-racist-tweets. Accessed: 2024-04-22.

[12] SoundGirls. Bitch boss vs boss bitch. https://soundgirls.org/bitch-boss-vs-boss-bitch/. Accessed: 2024-04-22.

[13] Northwest Asian Weekly. Racial slurs (chink). https://nwasianweekly.com/2011/09/racial-slurs-chink/. Accessed: 2024-04-22.

[14] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/N16-2013.

[15] Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752, 2019.

[16] Xin Shi and Yong Zheng. Perception and tolerance of sexual harassment: An examination of feminist identity, sexism, and gender roles in a sample of Chinese working women. *Psychology of Women Quarterly*, 44(2):217–233, 2020. URL https://doi.org/10.1177/0361684320903683.

[17] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360, 2020. ISSN 0306-4573. URL https://www.sciencedirect.com/science/article/pii/S0306457320308554.

[18] Peter Glick and Susan T Fiske. Ambivalent sexism. In *Advances in experimental social psychology*, volume 33, pages 115–188. Elsevier, 2001.

[19] Amber Wardell. Hostile and benevolent sexism: Two sides of the same coin. https://www.psychologytoday.com/us/blog/compassionate-feminism/202404/hostile-and-benevolent-sexism-two-sides-of-the-same-coin. Accessed: 2024-04-22.

[20] Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and*

Media, 27:100182, 2022. ISSN 2468-6964. URL https://www.sciencedirect.com/science/article/pii/S2468696421000604.

[21] Kate Manne. *Down girl: The logic of misogyny*. Oxford University Press, 2017.

[22] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer, 2018.

[23] Louise Richardson-Self. Woman-hating: On misogyny, sexism, and hate speech. *Hypatia*, 33(2):256–272, 2018.

[24] Jesse Fox, Carlos Cruz, and Ji Young Lee. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in human behavior*, 52:436–442, 2015.

[25] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 27(1):17–43, 2021.

[26] Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.woah-1.6.

[27] Matúš Pikuliak, Marián Šimko, and Mária Bieliková. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765, 2021. ISSN 0957-4174. URL https://www.sciencedirect.com/science/article/pii/S0957417420305893.

[28] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

[29] Aiqi Jiang and Arkaitz Zubiaga. Cross-lingual capsule network for hate speech detection in social media. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 217–223, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385510. URL https://doi.org/10.1145/3465336.3475102.

[30] Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*, 101:108005, 2022. ISSN 0045-7906. URL https://www.sciencedirect.com/science/article/pii/S0045790622002725.

[31] Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.c3nlp-1.2.

[32] Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Emotionally-bridged cross-lingual meta-learning for chinese sexism detection. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 627–639, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44696-2.

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

In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.

[34] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[36] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.747.

[37] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and*

Harms, pages 34–43, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.alw-1.5.

[38] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N15-1184.

[39] Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017. URL https://aclanthology.org/Q17-1022.

[40] Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-2051.

[41] Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. Mind your language: Abuse and offense detection for code-switched languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9951–9952, Jul. 2019. doi: 10.1609/aaai.v33i01.33019951. URL https://ojs.aaai.org/index.php/AAAI/article/view/5112.

[42] Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. Multilingual and multitarget hate speech detection in Tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II: Articles courts*, pages 351–360, Toulouse, France, 7 2019. ATALA. URL https://aclanthology.org/2019.jeptalnrecital-court.21.

[43] Muhammad Okky Ibrohim and Indra Budi. Translated vs non-translated method for multilingual hate speech identification in Twitter. *International Journal on Advanced Science, Engineering and Information Technology*, 9(4): 1116–1123, January 2019. ISSN 2088-5334. doi: 10.18517/ijaseit.9.4.8123.

[44] Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-3504.

[45] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[46] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. Overview of the Evalita 2018 task on Automatic Misogyny Identification (AMI). In *EVALITA@CLiC-it*, 2018.

[47] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. AMI@Evalita2020: Automatic misogyny identification. *Proceedings of the 7th evaluation campaign*

of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org, 2020.

[48] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.175.

[49] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[50] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[51] Kevin Michael DeLuca, Elizabeth Brunner, and Ye Sun. Weibo, WeChat, and the transformative events of environmental activism on China's wild public screens. *International Journal of Communication*, 10, 2016.

[52] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

[53] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*,

pages 7–16, Vancouver, Canada, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-2902.

[54] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 8:219563–219576, 2020.

[55] Marlis Hellinger and Anne Pauwels. 21. language and sexism. In *Handbook of language and communication: Diversity and change*, pages 651–684. De Gruyter Mouton, 2008.

[56] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.373.

[57] Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1174.

[58] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference*

on Web and Social Media, 15(1):573–584, 2021. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18085.

[59] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478, sep 2021. ISSN 1076-9757. URL https://doi.org/10.1613/jair.1.12590.

[60] Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Ratn Shah, and Debanjan Mahata. #YouToo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537, Florence, Italy, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1241.

[61] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the task on automatic misogyny identification at IberEval 2018. In *IberEval@SEPLN*, pages 214–228, 2018.

[62] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/S19-2007.

[63] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Devel-

oping a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France, 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL https://www.aclweb.org/anthology/2020.trac-1.25.

[64] Hala Mulki and Bilal Ghanem. Let-Mi: An Arabic Levantine Twitter dataset for misogynistic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv, Ukraine (Virtual), 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.wanlp-1.16.

[65] Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online, 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.114.

[66] Philine Zeinert, Nanna Inie, and Leon Derczynski. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online, 2021. Association for Computational Linguistics.

[67] Xiaochuang Han and Yulia Tsvetkov. Fortifying toxic speech detectors against disguised toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, 2020.

[68] Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021.

ISSN 0306-4573. URL https://www.sciencedirect.com/science/article/pii/S0306457321000339.

[69] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-1101.

[70] Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, and Ranka Stanković. Using lexical resources for irony and sarcasm classification. In *Proceedings of the 8th Balkan Conference in Informatics*, pages 1–8, 2017.

[71] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5(1): 11, 2016.

[72] Christopher Tuckwood. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www.hatebase.org*, 2017.

[73] Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS, 2018.

[74] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584, 2022. ISSN 0306-4379. URL https://www.sciencedirect.com/science/article/pii/S0306437920300715.

[75] Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.

[76] Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W16-5618.

[77] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-3504.

[78] Nabi Rezvani, Amin Beheshti, and Alireza Tabebordbar. Linking textual and contextual features for intelligent cyberbullying detection in social media. In *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, pages 3–10, 2020.

[79] Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.114.

[80] Angelo Basile and Chiara Rubagotti. CrotoneMilano for AMI at Evalita2018. a performant, cross-lingual misogyny detection system. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:206, 2018.

[81] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *The 6th In-*

ternational Conference on Learning Representations (ICLR), 2018. URL https://openreview.net/forum?id=H196sainb.

[82] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1474.

[83] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.470.

[84] Goran Glavaš, Mladen Karan, and Ivan Vulić. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.559.

[85] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany, 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-1133.

[86] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. Emoji-powered representation learning for cross-lingual sentiment classification. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference (WWW)*, pages 251–262. ACM, 2019. URL https://doi.org/10.1145/3308558.3313600.

[87] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335, 2016.

[88] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67:195–207, 2021.

[89] Aymé Arango, Jorge Pérez, and Barbara Poblete. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *arXiv preprint arXiv:2104.14728*, 2021.

[90] Satyajit Kamble and Aditya Joshi. Hate speech detection from code-mixed Hindi-English tweets using deep learning models. *arXiv preprint arXiv:1811.05145*, 2018.

[91] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access*, 9:106363–106374, 2021. doi: 10.1109/ACCESS.2021.3100435.

[92] Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 516–527, June 2018. URL https://aclanthology.org/N18-1048.

[93] Goran Glavaš and Ivan Vulić. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45, July 2018. URL https://aclanthology.org/P18-1004.

[94] Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, October-November 2018. URL https://aclanthology.org/D18-1026.

[95] Pedro Colon-Hernandez, Yida Xin, Henry Lieberman, Catherine Havasi, Cynthia Breazeal, and Peter Chin. RetroGAN: A cyclic post-specialization system for improving out-of-knowledge and rare word representations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2086–2095, August 2021. URL https://aclanthology.org/2021.findings-acl.183.

[96] Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, July 2019. URL https://aclanthology.org/P19-1070.

[97] Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2206–2217, November 2019. URL https://aclanthology.org/D19-1226.

[98] Dong Wang and Thomas Fang Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE, 2015.

[99] Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.383.

[100] Fedor Vitiugin, Yasas Senarath, and Hemant Purohit. Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback. In *Proceedings of the 13th ACM Web Science Conference*, WebSci '21, page 130–138, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383301. URL https://doi.org/10.1145/3447535.3462495.

[101] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

[102] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the GermEval 2018 shared task on the identification of offensive language. 2018.

[103] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg, pages 352–363, München [u.a.], 2019. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg. URL https://nbn-resolving.org/urn:nbn:de:bsz:mh39-93197.

[104] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/S19-2010.

[105] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics. URL https://aclanthology.org/2020.semeval-1.188.

[106] Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. Overview of OSACT4 Arabic offensive language detection shared

task. In Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak, editors, *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL https://aclanthology.org/2020.osact-1.7.

[107] Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, Tesconi Maurizio, et al. Overview of the Evalita 2018 hate speech detection task. In *CEUR Workshop Proceedings*, volume 2263, pages 1–9. CEUR, 2018.

[108] Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. HaSpeeDe2@Evalita2020: Overview of the Evalita 2020 hate speech detection task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, 2020.

[109] Mirko Lai, Fabio Celli, Alan Ramponi, Sara Tonelli, Cristina Bosco, and Viviana Patti. HaSpeeDe3 at Evalita 2023: Overview of the political and religious hate speech detection task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*, 2023.

[110] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450377508. URL https://doi.org/10.1145/3368567.3368584.

[111] Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '20, page 29–32, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389785. URL https://doi.org/10.1145/3441501.3441517.

[112] Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages and conversational hate speech. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 1–3, 2021.

[113] Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. Overview of the HASOC subtrack at FIRE 2022: Hate speech and offensive content identification in English and Indo-Aryan languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 4–7, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700231. URL https://doi.org/10.1145/3574318.3574326.

[114] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1226.

[115] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 1–5, 2020.

[116] Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. Results of the PolEval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter. 2019.

[117] Maaz Amjad, Alisa Zhila, Grigori Sidorov, Andrey Labunets, Sabur Butt, Hamza Imam Amjad, Oxana Vitman, and Alexander Gelbukh. UrduThreat@FIRE2021: Shared track on abusive threat identification in Urdu. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 9–11, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450395960. URL https://doi.org/10.1145/3503162.3505241.

[118] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of EXIST 2022: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 69:229–240, 2022.

[119] Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of EXIST 2023: sEXism Identification in Social neTworks. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 593–599. Springer, 2023.

[120] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech de-

tection. *Information Processing & Management*, 58(4):102544, 2021. ISSN 0306-4573. URL https://www.sciencedirect.com/science/article/pii/S0306457321000510.

[121] Bilingual Word Embeddings Rivaling. LLOD-driven bilingual word embeddings rivaling cross-lingual transformers in quality of life concept detection from French online health communities. In *Further with Knowledge Graphs: Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands*, volume 53, page 89. IOS Press, 2021.

[122] Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.woah-1.10.

[123] Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation*, pages 1–32, 2023.

[124] Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. Offensive language identification in low-resourced code-mixed Dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2108.12177*, 2021.

[125] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association*

for Computational Linguistics, 8:726–742, 2020. URL https://aclanthology.org/2020.tacl-1.47.

[126] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.41.

[127] Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2028, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://aclanthology.org/P16-1190.

[128] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1550.

[129] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

[130] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. 2018.

URL https://openreview.net/pdf?id=rkYTTf-AZ.

[131] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. URL https://aclanthology.org/Q19-1038.

[132] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.62.

[133] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1202.

[134] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Hybrid emoji-based masked language models for zero-shot abusive language detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 943–949, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.84.

[135] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in repre-*

sentation learning, ICML, volume 3, page 896. Atlanta, 2013.

[136] Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.ltedi-1.3.

[137] Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. Addressing the challenges of cross-lingual hate speech detection. *arXiv preprint arXiv:2201.05922*, 2022.

[138] Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. Improving zero-shot cross-lingual hate speech detection with pseudo-label finetuning of transformer language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1435–1439, May 2022. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19402.

[139] Zakaria Boulouard, Mariya Ouaissa, Mariyam Ouaissa, Moez Krichen, Mutiq Almutiq, and Karim Gasmi. Detecting hateful and offensive speech in Arabic social media using transfer learning. *Applied Sciences*, 12(24), 2022. ISSN 2076-3417. URL https://www.mdpi.com/2076-3417/12/24/12823.

[140] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*, 2020.

[141] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. A deep dive into multilingual hate speech classification. In Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science*

and Demo Track, pages 423–439, Cham, 2021. Springer International Publishing.

[142] Guizhe Song, Degen Huang, and Zhifeng Xiao. A study of multilingual toxic text detection approaches under imbalanced sample distribution. *Information*, 12(5), 2021. ISSN 2078-2489. URL https://www.mdpi.com/2078-2489/12/5/205.

[143] Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. Offensive content detection via synthetic code-switched text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6617–6624, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.575.

[144] Hajung Sohn and Hyunju Lee. MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559, 2019. doi: 10.1109/ICDMW.2019.00084.

[145] Fatima zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University - Computer and Information Sciences*, 34(8, Part B):6048–6056, 2022. ISSN 1319-1578. URL https://www.sciencedirect.com/science/article/pii/S1319157821001804.

[146] Mithun Das, Somnath Banerjee, and Animesh Mukherjee. Data bootstrapping approaches to improve low resource abusive language detection for Indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 32–42, New York, NY, USA, 2022. Association for Com-

puting Machinery. ISBN 9781450392334. URL https://doi.org/10.1145/3511095.3531277.

[147] Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.3.

[148] Anastasia Ryzhova, Dmitry Devyatkin, Sergey Volkov, and Vladimir Budzko. Training multilingual and adversarial attack-robust models for hate detection on social media. *Procedia Computer Science*, 213:196–202, 2022. ISSN 1877-0509. URL https://www.sciencedirect.com/science/article/pii/S1877050922017471. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society.

[149] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

[150] Ben Burtenshaw and Mike Kestemont. A Dutch dataset for cross-lingual multilabel toxicity detection. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 75–79, Online (Virtual Mode), September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.bucc-1.10.

[151] Neha Deshpande, Nicholas Farris, and Vidhur Kumar. Highly generalizable models for multilingual hate speech detection. *arXiv preprint arXiv:2201.11294*, 2022.

[152] Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee. Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1018–1023, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.96.

[153] Sayanta Paul, Sriparna Saha, and Jyoti Prakash Singh. COVID-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. *Multimedia tools and applications*, 82(6):8773–8789, 2023.

[154] Gretel Liz De la Peña Sarracén and Paolo Rosso. Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2196–2204, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.236.

[155] Sebastián E. Rodríguez, Héctor Allende-Cid, and Héctor Allende. Detecting hate speech in cross-lingual and multi-lingual settings using language agnostic representations. In João Manuel R. S. Tavares, João Paulo Papa, and Manuel González Hidalgo, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 77–87, Cham, 2021. Springer International Publishing. ISBN 978-3-030-93420-0.

[156] Lukas Stappen, Fabian Brunn, and Björn Schuller. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850*, 2020.

[157] Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.hackashop-1.5.

[158] Kartikey Pant and Tanvi Dadu. Towards code-switched classification exploiting constituent language resources. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 37–43, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-srw.6.

[159] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A corpus of English-Hindi code-mixed tweets for sarcasm detection. *arXiv preprint arXiv:1805.11869*, 2018.

[160] Vanessa Hahn, Dana Ruiter, Thomas Kleinbauer, and Dietrich Klakow. Modeling profanity and hate speech in social media with semantic subspaces. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*, pages 6–16, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.woah-1.2.

[161] Aiqi Jiang and Arkaitz Zubiaga. SexWEs: Domain-aware word embeddings via cross-lingual semantic specialisation for Chinese sexism detection in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):447–458, Jun. 2023. URL https://ojs.aaai.org/index.php/ICWSM/article/view/22159.

[162] Akshi Kumar and Nitin Sachdeva. Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. *Multimedia Systems*, 28(6):2027–2041, 2022.

[163] Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wassa-1.16.

[164] Anderson Almeida Firmino, Cláudio Souza de Baptista, and Anselmo Cardoso de Paiva. Using cross lingual learning for detecting hate speech in Portuguese. In Christine Strauss, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil, editors, *Database and Expert Systems Applications*, pages 170–175, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86475-0.

[165] Teodor Tiţa and Arkaitz Zubiaga. Cross-lingual hate speech detection using transformer models. *arXiv preprint arXiv:2111.00981*, 2021.

[166] Syrielle Montariol, Arij Riabi, and Djamé Seddah. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only, November 2022.

Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.33.

[167] Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management*, 59 (4):102981, 2022. ISSN 0306-4573. URL https://www.sciencedirect.com/science/article/pii/S0306457322000978.

[168] Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. Multilingual abusive comment detection at scale for Indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191, 2022.

[169] Tharindu Ranasinghe and Marcos Zampieri. An evaluation of multilingual offensive language identification methods for the languages of India. *Information*, 12(8), 2021. ISSN 2078-2489. URL https://www.mdpi.com/2078-2489/12/8/306.

[170] Tharindu Ranasinghe and Marcos Zampieri. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-demos.17.

[171] Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. Cross-lingual offensive language identification for low resource languages: The case of Marathi. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP*

2021), pages 437–443, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.50.

[172] Arianna Muti and Alberto Barrón-Cedeño. A checkpoint on multilingual misogyny identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-srw.37.

[173] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1), nov 2021. ISSN 2375-4699. URL https://doi.org/10.1145/3457610.

[174] Tung Minh Phung and Jan Cloos. An exploratory experiment on Hindi, Bengali hate-speech detection and transfer learning using neural networks. *arXiv preprint arXiv:2201.01997*, 2022.

[175] Neeraj Vashistha and Arkaitz Zubiaga. Online multilingual hate speech detection: Experimenting with Hindi and English social media. *Information*, 12(1), 2021. ISSN 2078-2489. URL https://www.mdpi.com/2078-2489/12/1/5.

[176] Aya Elouali, Zakaria Elberrichi, and Nadia Elouali. Hate speech detection on multilingual Twitter using convolutional neural networks. *Rev. d'Intelligence Artif.*, 34(1):81–88, 2020.

[177] Adeep Hande, Siddhanth U Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced Dravidian languages. *arXiv preprint arXiv:2108.03867*, 2021.

[178] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896, 2022. doi: 10.1109/ACCESS.2022.3147588.

[179] Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, pages 1–10, 2023. doi: 10.1109/TCSS.2023.3252401.

[180] Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559, 2021.

[181] Fatemah Husain and Ozlem Uzuner. Transfer learning across Arabic dialects for offensive language detection. In *2022 International Conference on Asian Language Processing (IALP)*, pages 196–205, 2022. doi: 10.1109/IALP57159.2022.9961263.

[182] Chanhee Lee, Kisu Yang, Taesun Whang, Chanjun Park, Andrew Matteson, and Heuiseok Lim. Exploring the data efficiency of cross-lingual post-training in pretrained language models. *Applied Sciences*, 11(5), 2021. ISSN 2076-3417. URL https://www.mdpi.com/2076-3417/11/5/1974.

[183] Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.woah-1.15.

[184] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[185] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[186] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online (v2020.3)*. Zenodo, 2013. URL https://doi.org/10.5281/zenodo.7385533.

[187] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/N16-3020.

[188] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[189] Wikipedia. Sina Weibo — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Sina_Weibo. Accessed: 2021-01-03.

[190] SinaFinance. Sina weibo monthly active users reach 550 million, revenue exceeds wall street expectations. https://finance.sina.com.cn/stock/usstock/c/2020-05-19/doc-iircuyvi3963989.shtml. Accessed: 2021-01.

[191] Bernardo Huberman, Daniel M Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 2009.

[192] Qing Xu, Ziyi Shen, Neal Shah, Raphael Cuomo, Mingxiang Cai, Matthew Brown, Jiawei Li, and Tim Mackey. Characterizing Weibo social media posts from Wuhan, China during the early stages of the COVID-19 pandemic: Qualitative content analysis. *JMIR Public Health and Surveillance*, 6(4):e24125, 2020.

[193] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[194] Hsu Yang and Chuan-Jie Lin. TOCP: A dataset for Chinese profanity processing. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 6–12, Marseille, France, 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL https://www.aclweb.org/anthology/2020.trac-1.2.

[195] Leonardo Betti, Chiara Abrate, and Andreas Kaltenbrunner. Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(10), 2023. doi: 10.1140/epjds/s13688-023-00384-8. URL https://doi.org/10.1140/epjds/s13688-023-00384-8.

[196] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[197] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Pro-*

cessing (EMNLP), pages 1746–1751, Doha, Qatar, 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D14-1181.

[198] Li-Ping Jing, Hou-Kuan Huang, and Hong-Bo Shi. Improved feature selection approach TFIDF in text mining. In *Proceedings of International Conference on Machine Learning and Cybernetics*, volume 2, pages 944–946. IEEE, 2002.

[199] Manuela Thomae and Afroditi Pina. Sexist humor and social identity: the role of sexist humor in men's in-group cohesion, sexual harassment, rape proclivity, and victim blame. *HUMOR*, 28(2):187–204, 2015. URL https://doi.org/10.1515/humor-2015-0023.

[200] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[201] Enyan Dai, Yiwei Sun, and Suhang Wang. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862, 2020.

[202] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2cad8fa47bbef282badbb8de5374b894-Abstract.html.

[203] Geoffrey E. Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HJWLfGWRb.

[204] Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. Investigating capsule networks with dynamic routing for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3110–3119, Brussels, Belgium, 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1350.

[205] Saurabh Srivastava, Prerna Khurana, and Vartika Tewari. Identifying aggression and toxicity in comments using capsule network. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 98–105, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4412.

[206] Saurabh Srivastava and Prerna Khurana. Detecting aggression and toxicity using a multi dimension capsule network. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 157–162, Florence, Italy, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-3517.

[207] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1299.

[208] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[209] Erik Cambria, Robyn Speer, Catherine Havasi, and Amir Hussain. SenticNet: A publicly available semantic resource for opinion mining. In *2010 AAAI Fall*

Symposium Series, 2010.

[210] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. Analyzing the hate and counter speech accounts on Twitter. *arXiv preprint arXiv:1812.02712*, 2018.

[211] Yanqing Chen and Steven Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland, 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P14-2063.

[212] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[213] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D14-1179.

[214] Manuel Montes, Paolo Rosso, Julio Gonzalo, Ezra Aragón, Rodrigo Agerri, Miguel Ángel Álvarez-Carmona, Elena Álvarez-Mellado, Jorge Carrillo-de Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez Zafra, Salvador Lima, Flor Miriam Plaza-de Arco, and Mariona Taulé. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*, 2021.

[215] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting*

of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy, 2019. Association for Computational Linguistics.

[216] Tianyu Zhang and Fucheng You. Research on short text classification based on TextCNN. In *Journal of Physics: Conference Series*, volume 1757, page 012092. IOP Publishing, 2021.

[217] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537, 2011.

[218] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1095.

[219] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, 2019. Association for Computational Linguistics.

[220] Yoav Goldberg and Omer Levy. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[221] Nikolaos Pappas and James Henderson. GILE: A generalized input-label embedding for text classification. *Transactions of the Association for Com-*

putational Linguistics (TACL), 7(0):139–155, 2019. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1550.

[222] Xiao Li, Jiaxing Song, and Weidong Liu. Label-attentive hierarchical attention network for text classification. In *Proceedings of the 2020 5th International Conference on Big Data and Computing*, ICBDC 2020, page 90–96, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375474. URL https://doi.org/10.1145/3404687.3404706.

[223] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, 2010.

[224] Goran Glavaš and Ivan Vulić. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 181–187, June 2018. URL https://aclanthology.org/N18-2029.

[225] Haozhou Wang, James Henderson, and Paola Merlo. Multi-adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 463–472, June 2021. URL https://aclanthology.org/2021.naacl-main.39.

[226] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium, October-November

2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1330.

[227] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[228] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[229] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, June 2013. URL https://aclanthology.org/N13-1092.

[230] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.

[231] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980.

[232] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, jul 2011. ISSN 1532-4435. URL http://jmlr.org/papers/v12/duchi11a.html.

[233] Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4): 665–695, 2015.

[234] Peng Jin and Yunfang Wu. SemEval-2012 task 4: Evaluating Chinese word similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, page 374–377, 2012. URL https://aclanthology.org/S12-1049.

[235] Xiang Wang, Yan Jia, Bin Zhou, Zhao-Yun Ding, and Zheng Liang. Computing semantic relatedness using Chinese Wikipedia links and taxonomy. *Journal of Chinese Computer Systems*, 32(11):2237–2242, 2011.

[236] Tzu-ray Su and Hung-yi Lee. Learning Chinese word representations from glyphs of characters. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 264–273, September 2017. URL https://www.aclweb.org/anthology/D17-1025.

[237] Jerrold H Zar. Spearman rank correlation: overview. *Wiley StatsRef: Statistics Reference Online*, 2014.

[238] Chi Sun, Xipeng Qiu, and Xuanjing Huang. VCWE: Visual character-enhanced word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2710–2719, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1277.

[239] Zongyang Xiong, Ke Qin, Haobo Yang, and Guangchun Luo. Learning Chinese word representation better by cascade morphological n-gram. *Neural Computing and Applications*, 33(8):3757–3768, 2021.

[240] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.58.

[241] Georgios Rizos, Konstantin Hemker, and Björn Schuller. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, page 991–1000. ACM, 2019. URL https://doi.org/10.1145/3357384.3358040.

[242] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[243] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclanthology.org/D16-1229.

[244] Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. Resources for multilingual hate speech detection. In *Pro-*

ceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 122–130, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.woah-1.12.

[245] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[246] OpenAI. GPT-4 technical report, 2023.

[247] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[248] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[249] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[250] Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. AdaPrompt: Adaptive model training for prompt-based NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6057–6068, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.448.

[251] Lawrence Han and Hao Tang. Designing of prompts for hate speech recognition with in-context learning. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–320. IEEE, 2022.

[252] Sarthak Roy, Ashish Harshavardhan, Animesh Mukherjee, and Punyajoy Saha. Probing LLMs for hate speech detection: strengths and vulnerabilities. *arXiv preprint arXiv:2310.12860*, 2023.

[253] Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. Interpretable unified language checking. *arXiv preprint arXiv:2304.03728*, 2023.

[254] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.234.

[255] Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6255–6263. International Joint Conferences on Artificial Intelligence Organization, 8 2023. URL https://doi.org/10.24963/ijcai.2023/694.

[256] Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates, December

2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.744.

[257] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.796.

[258] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. DravidianCodeMix: Sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806, 2022.

[259] John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval)*, pages 59–69, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.semeval-1.6.

[260] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[261] Nauros Romim, M. Firoz Ahmed, Hriteshwar Talukder, and Md. Saiful Islam. Hate speech detection in the Bengali language: A dataset and its baseline eval-

uation. *ArXiv*, abs/2012.09686, 2020. URL https://api.semanticscholar.org/CorpusID:229298046.

[262] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*, 2020.

[263] Muhammad Moin Khan, Khurram Shahzad, and Muhammad Kamran Malik. Hate speech detection in Roman Urdu. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1), mar 2021. ISSN 2375-4699. URL https://doi.org/10.1145/3414524.

[264] Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wanlp-1.13.

[265] Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.629.

[266] Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17:100071, 2020.

[267] Ravi Shekhar, Marko Pranjic, Senja Pollak, Andraz Pelicon, and Matthew Purver. Automating news comment moderation with limited resources: Bench-

marking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics*, 2020.

[268] Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, and Muhammad Tariq Sadiq. Automatic detection of offensive language for Urdu and Roman Urdu. *IEEE Access*, 8:91213–91226, 2020.

[269] Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2512–2522, 2020.

[270] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*, 2020.

[271] Çağrı Çöltekin. A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, 2020.

[272] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.430.

[273] Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*,

pages 54–63, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.peoples-1.6.

[274] Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, pages 103–114. Springer, 2019.

[275] Masaki Arata. *Study on change of detection accuracy over time in cyberbullying detection*. PhD thesis, Master's thesis, Kitami Institute of Technology, Department of Computer Science, 2019.

[276] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. URL https://doi.org/10.1145/3331184.3331262.

[277] Muhammad Okky Ibrohim and Indra Budi. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-3506.

[278] Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-3512.

[279] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in Twitter. *Sensors*, 19(21), 2019. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/19/21/4654.

[280] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-3510.

[281] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1271.

[282] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5102.

[283] Michał Ptaszyński, Gniewosz Leliwa, Mateusz Piech, and Aleksander Smywiński-Pohl. Cyberbullying detection–technical report 2/2018, department of computer science AGH, university of science and technology. *arXiv preprint arXiv:1808.00926*, 2018.

[284] Azalden Alakrot, Liam Murray, and Nikola S Nikolov. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Computer Science*, 142:174–181, 2018.

[285] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-1105.

[286] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[287] Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5118.

[288] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1443.

[289] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and

Nicolas Kourtellis. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[290] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348966. URL https://doi.org/10.1145/3091478.3091509.

[291] Dmitry Devyatkin, Ivan Smirnov, Margarita Ananyeva, Maria Kobozeva, Andrey Chepovskiy, and Fyodor Solovyev. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts). In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 188–190, 2017. doi: 10.1109/ISI.2017.8004907.

[292] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.

[293] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[294] Rogers de Pelle and Viviane Moreira. Offensive comments in the Brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil, 2017. SBC. URL https://sol.sbc.org.br/index.php/brasnam/article/view/3260.

[295] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[296] Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238, 2017. doi: 10.1109/ICACSIS.2017.8355039.

[297] Paula Fortuna. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. 2017. URL https://api.semanticscholar.org/CorpusID:65158765.

[298] Uwe Bretschneider and Ralf Peters. Detecting offensive statements towards foreigners in social media. 2017.

[299] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-3008.

[300] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Automatic detection and

prevention of cyberbullying. In *International Conference on Human and Social Analytics (HUSO 2015)*, pages 13–18. IARIA, 2015.

[301] Michal Ptaszynski1 Pawel Dybala, Tatsuaki Matsuba2 Fumito Masui, Rafal Rzepka, and Kenji Araki. Machine learning and affect analysis against cyberbullying. In *Proceedings of the Linguistic and Cognitive Approaches to Dialog Agents Symposium*, 2010.

A

Overview Tables of Cross-lingual Hate Speech Studies

We present two comprehensive tables in this appendix, containing multilingual datasets and CLTL techniques used in surveyed papers respectively.

A.1 Summary of Multilingual Data Resources

Ref	Year	Topic	Source	Language	#Sample	Label	#Cit	CM?	Avail?
[153]	2023	cyberbullying	Twitter	en, hi	22k	1	<50	✓	✗
[256]	2022	offence	NAVER-news, Youtube	ko	40,429	1,2,3	<10	✗	✓
[20]	2022	Sexism	Weibo	zh	8,969	1,2,3	<50	✗	✓
[143]	2022	offence	-	en, fr, de, es	37,221	1	<10	✓	✓
[257]	2022	offence	Weibo, Zhihu	zh	37,480	1,2	<50	✗	✓
[168]	2022	abuse	ShareChat	hi, kn, ml, ta, te	92,881	1	<10	✓	✓
[258]	2022	offence	Youtube	ta	60k	1,3	<50	✓	✓
[244]	2022	offence	Twitter	es	9,834	2	<10	✗	✓
[63]	2022	aggressiveness, misogyny	Youtube	bn	25k	1	<100	✗	✓
[162]	2022	cyberbullying	Facebook, Twitter	en, hi	6,500	1	<50	✓	✗
[259]	2021	toxicity	Civil Comments	en	10,629	1,2	<100	✗	✓
[150]	2021	toxicity	Ask.fm	nl	10,189	2	<10	✗	✗
[124]	2021	offence	Youtube	kn, ml, ta	71,691	1,3	<20	✓	✓
[260]	2021	hate speech	Gab, Twitter	en	20,148	1,2	<500	✗	✓
[261]	2021	hate speech	Facebook, Youtube	bn	30k	1	<50	✗	✗
[112]	2021	hate speech, offence	Twitter	en, hi, mr	13,755	1,4	<100	✗	✓
[117]	2021	abuse	Twitter	ur	8,400	1	<50	✗	✓
[262]	2021	hate speech	open-source platform	en	41,255	1,2	<100	✗	✓
[171]	2021	offence	Twitter	mr	2499	1,2,3	<50	✗	✓
[263]	2021	hate speech, offence	Twitter	roman ur	5k	1	<50	✗	✗
[264]	2021	offence	Twitter	ar	10k	1,2	<250	✗	✓
[265]	2020	offence	Twitter	el	4,779	1	<250	✗	✓
[266]	2020	hate speech	Twitter	en, fr, de, el, es	264,035	1	<50	✗	✓
[84]	2020	hate speech, abuse	Facebook, Fox news, Twitter, Wikipedia	sq, hr, en, de, ru, tr	109,955	1	<50	✗	✓
[267]	2020	offence	24sata, Eesti Ekspress, Vecernji List	hr, et	62.6m	-	<10	✗	✗
[268]	2020	offence, abuse	Youtube	Roman ur,	12,171	1	<100	✗	✓

[Continued table]

				ur					
[269]	2020	hate speech, offence	Twitter	Roman ur	10,012	1,2	<50	✗	✓
[111]	2020	offence	Twitter, Youtube	en, de, hi, ml, ta	15,047	1,2	<250	✓	✓
[108]	2020	hate speech	Twitter	it	12,081	1	<100	✗	✓
[270]	2020	hate speech	Gab	en	27,655	1	<250	✗	✓
[271]	2020	offence	Twitter	tr	36,232	1,3	<250	✗	✓
[115]	2020	aggressiveness	Youtube	bn, en, hi	15k	1	<250	✗	✓
[272]	2020	abuse, offence	Facebook, Reddit, Twitter	da	3,600	1,3	<250	✗	✓
[273]	2020	offence	Youtube	kn	7,671	1,3	<100	✓	✗
[105]	2020	offence	Twitter	ar, da, en, el, tr	9m	1,3	<500	✗	✓
[274]	2019	offence	Facebook	en, sl	22,877	1,2	<50	✓	✗
[275]	2019	cyberbullying	Twitter	ja	4,096	1	<10	✗	✗
[276]	2019	hate speech	Twitter	en	14,949	1	<250	✗	✓
[277]	2019	abuse, hate speech	Twitter	id	5,561	1,2,3	<250	✗	✓
[278]	2019	abuse, toxicity hate speech	Twitter	at	5,846	1	<250	✗	✗
[279]	2019	hate speech	Twitter	es	6k	1	<100	✗	✓
[280]	2019	hate speech, offence	Twitter	pt	5,668	1	<100	✗	✓
[101]	2019	cyberbullying, toxicity	Twitter	pl	11,041	1,2	<50	✗	✓
[103]	2019	offence	Twitter	de	7,025	1,2	<250	✗	✓
[62]	2019	aggressiveness	Twitter	en	19,600	1,3	<750	✗	✓
[104]	2019	offence	Twitter	en, es	14k	1,3	<700	✗	✓
[110]	2019	hate speech, offence	Facebook, Twitter	en, de, hi	17,657	1,2,3	<100	✗	✓
[281]	2019	hate speech	Twitter	en, fr, it	4,078	1,2	<250	✗	✓
[82]	2019	hate speech	Twitter	ar, en, fr	13,014	1,2,3	<250	✗	✓
[42]	2019	sexism, misogyny	Twitter	fr	3,085	1	<50	✗	✗
[107]	2018	hate speech	Facebook, Twitter	en, it	4k	1	<250	✗	✓
[282]	2018	hate speech	Stormfront	en	10,568	1	<500	✗	✓
[283]	2018	cyberbullying	Formspring.me	en	12,772	1	<50	✗	✗
[284]	2018	abuse, offence	Youtube	ar	167,549	1	<100	✗	✓
[285]	2018	hate speech	Twitter	en, hi	4,575	1	<250	✓	✓
[286]	2018	hate speech	Twitter	en	27,330	1,2	<250	✗	✓
[46]	2018	misogyny	Twitter	en, it	20k	1,2,3	<250	✗	✓
[287]	2018	abuse, offence, hate speech	Twitter	en, hi	17,698	1,2	<100	✗	✓
[114]	2018	aggressiveness	Facebook, Twitter	en, hi	39k	1,2,3	<250	✓	✓
[44]	2018	offence	Twitter	en, hi	3,679	1,2	<250	✓	✗

[**Continued table**]

[288]	2018	aggressiveness, hate speech, offence	Twitter	it	6k	2,4	<250	✗	✓
[289]	2018	abuse, hate speech	Twitter	en	80k	1,2	<550	✗	✓
[102]	2018	hate speech, offence	Twitter	de	8,541	1,2	<500	✗	✓
[61]	2018	misogyny	Twitter	en, es	8,115	1,2,3	<250	✗	✓
[290]	2017	harassment, offence, racism	BlockTogether, Twitter	en	35k	1	<100	✗	✓
[291]	2017	offence	Twitter	ru	493	1,2	<50	✗	✗
[292]	2017	hate speech	Fox news	en	1,528	1	<250	✗	✓
[293]	2017	hate speech	Wikipedia	en	115,737	1	<750	✗	✓
[294]	2017	hate speech, offence	g1.globo.com	pt	1,250	1	<100	✗	✓
[295]	2017	hate speech, offence	Twitter	de	541	1,2,4	<500	✗	✓
[296]	2017	hate speech	Twitter	id	713	1	<250	✗	✓
[297]	2017	hate speech	Twitter	pt	5,668	1,2	<50	✗	✓
[298]	2017	hate speech, offence	Facebook	de	5,836	1,3	<100	✗	✓
[49]	2017	offence	Twitter	en	24,802	1,2	<2200	✗	✓
[299]	2017	abuse, offence, hate speech	Twitter	ar	32k	1	<500	✗	✓
[14]	2016	hate speech	Twitter	en	16,914	1	<1550	✗	✓
[300]	2015	cyberbullying	Ask.fm	nl	85,462	2,4	<250	✗	✗
[301]	2010	cyberbullying	informal sites of Japanese secondary schools	ja	2,999	1	<100	✗	✗

Table A.1: Summary of included dataset resources in automated identification of cross-lingual hate speech phenomena and sorted by released year. Language names are represented by using the standardized nomenclature ISO 639-1. "Ref" = "reference", "#Sample" = "number of instances", "%Hate" = "percentage of hateful texts", "#Cit" = "number of citations", "CM?" = "whether or not the dataset is code-mixed", "Avail?" = "whether or not the dataset is available", and "Label Type" denotes the annotation scheme: (1) binary labels, (2) fine-grained category of offensive content, (3) attack target, and (4) intensity score.

A.2 Summary of Cross-lingual Techniques

Ref	Year	Transfer Level	Model	Approach	Avail?
[153]	2023	Feature	BERT, CNN, LR, LSTM, MLP, SVM	propose an ensemble model of deep neural networks (MLP, CNN, BiLSTM and BERT) with random and Xavier initialization of weights based on aligned word embeddings in source and target languages	✗
[179]	2023	Parameter	XLM-R, mBERT	propose HateMAML, a model-agnostic meta-learning-based framework that uses a semi-supervised self-refinement strategy to fine-tune a better pre-trained model for unseen data in target language	✗
[123]	2023	Feature, Parameter	CNN, LSTM, mBERT	propose an ensemble-based cross-lingual approach by leveraging cross-lingual word embeddings to train CNN/BiLSTM classifiers and directly train mBERT on English dataset, generating pseudo labels for two unlabelled German datasets by an ensemble of three trained models, and fine-tuning them on bootstrapping German datasets	✗
[161]	2023	Feature	BERT, CNN, MacBERT	propose a domain-aware cross-lingual semantic specialisation framework between source and target languages to construct sexism-specific word embeddings (SexWEs)	✓
[31]	2023	Instance, Parameter	XLM, mBERT, BERT, RoBERTa	Investigate the impact of cultural background differences based on Korean / English and Chinese languages in zero-/few-shot and translation settings	✗
[26]	2023	Parameter	BERT, DeBERTa, FLAN-T5, RoBERTa, XLM-R, mT0	explore different prompting formats on multiple hate speech datasets, and compare the zero-shot learning performance of encoder models with the recent LLMs based on instruction fine-tuning	✓
[162]	2022	Instance, Parameter	CapsNet, LSTM, MLP	propose MIIL-DNN, a multi-input integrative learning framework based on deep neural networks, combining information from three paralleled sub-networks by using model-level multi-lingual fusion strategy to detect English-Hindi code-mixed bully content	✗
[145]	2022	Instance, Parameter	AraBERT, BERT, mBERT	propose a joint learning framework by fine-tuning mBERT on mixed datasets, and translation-based methods using BERT and AraBERT for English and Arabic languages	✗
[146]	2022	Instance, Parameter	MuRIL, mBERT	perform a large-scale analysis of cross-lingual hate speech by investigating the performance of multilingual models (mBERT and MuRIL) on four different transfer strategies across eight different Indic languages	✗
[183]	2022	Parameter	XLM-T	propose Multilingual HateCheck (MHC), a suite of functional tests for multilingual hate speech detection models, and fine-tune multilingual XLM-T on individual datasets and combined datasets in Spanish, Italian and Portuguese	✓
[138]	2022	Instance, Parameter	BERT, RoBERTa, XLM-R	propose a cross-lingual transfer approach by training XLM-R in a zero-shot setting to generate pseudo-labels for target data, and then using it to fine-tune monolingual pre-trained models	✓

[Continued table]

[178]	2022	Parameter	XLM-R	propose a cross-lingual meta learning-based approach to fine-tuning the base learner XLM-R with parallel few-shot datasets in different target languages by using optimisation-based Model-Agnostic Meta-Learning (MAML) and Proto-MAML models	✗
[167]	2022	Parameter	XLM-R, mBERT	select optimal transfer languages based on the correlation between linguistic similarity and zero-shot cross-lingual performance of mBERT/XLM-R on 7 different languages, and propose a new linguistic similarity metric based on WALS	✗
[151]	2022	Instance, Parameter	CNN-GRU, LR, mBERT	propose multilingual experiments for a compiled dataset of 11 languages by training the model on all available languages or a particular language family, and testing it on each language (in the family)	✓
[244]	2022	-	BERT, CNN-GRU, LSTM, LR, XLM-R	perform a comparative study of existing cross-lingual architectures on multilingual datasets including self-created Spanish dataset	✓
[30]	2022	Instance, Feature, Parameter	LSTM	Propose a mapping method between source and target language BERT embeddings into a shared space using adversarial training and Procrustes analysis, and propose an agreement regularised training schema to select source data which is most similar to target one based on shared embeddings to fine-tune the trained LSTM model	✓
[154]	2022	Feature, Parameter	GNN, USE, XLM-R, mBERT	propose a Graph Auto-Encoders (GAE) framework to learn embeddings of a set of texts in an unsupervised way, and add prior language knowledge using Universal Sentences Encoder (USE) in a multilingual setting	✗
[99]	2022	Parameter	XLM-T	fine-tune pre-trained multilingual model XLM-T by using randomly sampled differently-sized datasets in target language	✓
[139]	2022	Instance, Parameter	AraBERT, BERT, mBERT	propose a selection of BERT-based models in a cross-lingual setting, covering texts written in the standard Arabic, as well as three of most spoken Arabic dialects in the region (Egyptian, Iraqi, and Gulf)	✗
[137]	2022	Feature, Parameter	CNN, LSTM, mBERT	propose an ensemble-based cross-lingual approach by leveraging cross-lingual word embeddings to train CNN/BiLSTM classifiers and directly train mBERT on English dataset, generating pseudo labels for two unlabelled German datasets by an ensemble of three trained models, and fine-tuning them on bootstrapping German datasets	✗
[174]	2022	Parameter	LSTM	propose a cross-lingual approach between Hindi and Bengali by reusing monolingual Hindi classifier without embedding layers and replacing embedding layer with untrained Bengali embedding layer to build Hindi-Bengali cross-lingual classifier	✗
[172]	2022	Parameter	BERT, mBERT	explore the feasibility of detecting misogyny through a transfer learning approach by fine-tuning mBERT in a zero-shot setting or on different combinations of multiple languages (English, Italian and Spanish)	✗
[148]	2022	Instance, Parameter	XLM-R	use multilingual source datasets and an adversarial attacked source dataset to consistently fine-tune an ensemble of XLM-R models and test it on the target dataset in a zero-shot way	✗

[Continued table]

[166]	2022	Parameter	XLM-R, XLM-T, mBERT	propose a zero-shot cross-lingual approach by applying XLM-R to a MACHAMP multi-task architecture to jointly train hate speech detection with auxiliary tasks	✓
[168]	2022	Parameter	AbuseXLMR, MuRIL, XLM-R	create a multilingual abusive dataset (MACD) and abuse-specific pre-trained model AbuseXLMR, and perform four cross-lingual strategies (zero-shot, few-shot, joint training, and pre-training) across five Indic languages	✓
[181]	2022	Parameter	AraBERT	explore cross-lingual performance across Arabic dialects (Levantine, Egyptian and Tunisian) by directly fine-tuning AraBERT or keep pre-training it on different combinations of dialects	✗
[143]	2022	Instance, Parameter	XLM-R	release a human-generated dataset for testing for three language combinations en-fr, en-es, and en-de and a synthetic code-switched dataset for multilingual training based on XLM-R model	✗
[141]	2021	Instance, Parameter	LR, mBERT	investigate cross-lingual zero-shot learning by using multi-source languages to train mBERT or LR with LASER embeddings and fine-tuning the trained model in incremental amounts of target data based on 9 languages	✓
[120]	2021	Instance, Parameter	LSTM, mBERT	propose a joint-learning cross-lingual approach to detect hate speech, encoding parallel source and target datasets via multilingual representations (MUSE and mBERT) and integrating multilingual hate speech lexicon features (HurtLex) together into LSTM networks	✗
[171]	2021	Parameter	XLM-R	create Marathi Offensive Language Dataset (MOLD), and propose zero-shot, few-shot and weight-frozen cross-lingual approaches based on XLM-R, using three source languages (English, Hindi and Bengali) to identify Marathi offensive content	✓
[79]	2021	Parameter	XLM-R, mBERT	explore the limits of cross-lingual hate speech detection based on four different monolingual and cross-lingual learning settings by fine-tuning mBERT/XLM-R	✗
[173]	2021	Parameter	XLM-R, mBERT	propose a cross-lingual approach to train XLM-R and mBERT classifier on source language (English) and save model weights to initialise the model for target language	✓
[175]	2021	Instance, Parameter	BERT, CNN-LSTM, LR	experimented with several models including LR, CNN-LSTM, and BERT to build a multilingual system trained on code-switched datasets in English and Hindi by adopting a transfer learning approach	✓
[170]	2021	Parameter	XLM-R	propose a multilingual framework MUDES based on XLM-R, and fine-tune MUDES on English source data and evaluate the model on two target datasets in Danish and Greek	✓
[124]	2021	Instance, Parameter	DistilmBERT, IndicBERT, MuRIL, ULMFiT, XLM-R, mBERT	construct a transliterated dataset based on code-mixed texts in Kannada, Malayalam and Tamil with pseudo-labels generated by BERT-based models, and fine-tune the pre-trained model on both newly constructed and code-mixed datasets	✓

[Continued table]

[177]	2021	Parameter	ALBERT, BERT, CharacterBERT, DistilBERT, RoBERTa, XLM, XLM-R, XLNet	propose a multi-task learning approach based on sentiment analysis and offensive language identification tasks using diverse pre-trained multilingual models	✓
[169]	2021	Parameter	XLM-R	propose different cross-lingual approaches in zero-shot, few-shot and multi-source training settings across six languages from the two most widely spoken language families in India	✗
[136]	2021	Feature, Parameter	CNN, LSTM, SVM	propose an ensemble-based cross-lingual approach by leveraging bilingual word embeddings to train four neural classifiers (CNN/BiLSTM) on English dataset, generating pseudo labels for two unlabelled German datasets by an ensemble of four trained models, and fine-tuning them on bootstrapping German datasets	✗
[157]	2021	Feature, Parameter	MLP, mBERT	propose a zero-shot cross-lingual transfer approach by training mBERT or LASER embeddings in multilayer perceptron classifier on English dataset and evaluate models on five different target languages	✓
[180]	2021	Parameter	cseBERT, mBERT	propose cross-lingual intermediate training regimes by training mBERT / cseBERT on one or more non-target languages (English, Slovenian and Arabic) and then fine-tuning the trained model on different amounts (from 0 to 100%) of the five target languages	✓
[142]	2021	Instance, Parameter	XLM-R, mBERT	propose an ensemble strategy to combine different loss functions (BCE and Focal) and multiple pre-trained models (mBERT and XLM-R) into nice combination sets using macro F1 scores as the fusion weights	✗
[100]	2021	Feature	LSTM	propose a Multilingual Interactive Attention Network (MLIAN) model by building upon frame semantics theory with attention weights for interpretability and human-in-the-loop paradigm for model adaptability on multilingual corpora	✗
[155]	2021	Feature, Parameter	DT, LabSE, RF, SVM, mBERT	propose zero-shot cross-lingual learning approaches by training SMV-based and tree-based classifiers with LabSE and mBERT Embeddings, or directly training LabSE and mBERT models	✗
[29]	2021	Instance, Parameter	CNN, MLP, CapsNet, LSTM, SVM, XLM-R, mBERT	propose a cross-lingual capsule network learning model (CCNL-Ex) by infusing extra hate speech lexical semantics into parallel embeddings of source and translated target data, and then connect them to capsule networks for detection	✗
[160]	2021	Feature	LDA	propose an approach to learn semantic sub-spaces to model profane language on both word and sentence level representations and evaluate their generalisability on a variety of similar and distant target languages in a zero-shot cross-lingual setting	✓
[165]	2021	Parameter	XLM-R, mBERT	experiment with fine-tuned altered versions of mBERT and XLM-R to adopt cross-lingual transfer learning on English data for training and French data for testing	✗

[Continued table]

[89]	2021	Feature	BERT	propose a hate-specific data representation by constructing monolingual vector spaces and utilising bilingual dictionaries for alignment	✗
[164]	2021	Parameter	XLM-R	explore three cross-lingual learning methods (Zero-Shot Transfer, Joint Learning and Cascade Learning) based on XLM-R by using Italian as the source language and Portuguese as the target language	✗
[150]	2021	Parameter	LSTM, RF, mBERT	propose a random forest ensemble of LSTM using MUSE embeddings and mBERT model on a cross-validated training set with grid-searched parameters	✗
[163]	2021	Parameter	BERT, CNN, LSTM	compare the performance of stylometric and emotion-based features with commonly used features and SOTA deep learning models on multilingual hate speech datasets	✗
[140]	2020	Instance, Feature, Parameter	BERT, LR, mBERT	analyse multilingual hate speech in 9 languages from 16 different sources, and conduct experiments in both monolingual and multilingual settings	✓
[83]	2020	Parameter	XLM-R	propose a cross-lingual approach to train XLM-R classifier on source language (English) and save model weights to initialise the model for target languages	✓
[17]	2020	Instance, Parameter	BERT, SVM, LSTM	present a comparative analysis of different cross-lingual models, such as momo-lingual deep learning models with translation, and joint models with multilingual embeddings and lexical hate features	✓
[84]	2020	Parameter	XLM-R, mBERT	create a multi-domain and multilingual evaluation dataset (XHate-999) for abusive language detection, and propose a zero-shot cross-lingual approach and a cross-lingual adaptation via intermediate masked language modelling on filtered target data	✓
[156]	2020	Parameter	LSTM, XLM	propose a cross-lingual architecture of using frozen Transformer Language Model (TLM) as the encoder with Attention-Maximum-Average Pooling (AXEL) in zero-shot and few-shot settings	✗
[74]	2020	Instance, Parameter	DT, LSTM	translate Spanish dataset to the English language, and use multilingual word embedding representations MUSE in a Gradient Boosted Decision Tree and an LSTM-based model	✓
[176]	2020	Instance, Parameter	CNN	propose a CNN-based architecture with character level representations and combine datasets in different languages into two versions of multilingual datasets for training and testing	✗
[152]	2020	Feature	HAN	experiment with Hierarchical Attention Network (HAN) by concatenating switching pattern features between Hindi and English into the last hidden layer of HAN	✗
[134]	2020	Feature	XLM	propose a Hybrid Emoji-based Masked Language Model (HE-MLM) to leverage the common information conveyed by emojis across different languages to improve the learned cross-lingual representations of social media texts in a zero-shot setting	✗
[158]	2020	Instance, Parameter	RoBERTa, ULMFiT, XLM-R	Convert English-Hindi code-mixed data into the high resource languages (English) with translation and transliteration in a cross-lingual setting based on XLM-R model	✗
[82]	2019	Parameter	LSTM, Sluice NNs	propose a multitask learning architecture based on Sluice Networks coupled with Babylon and MUSE embeddings	✗

[**Continued table**]

[40]	2019	Instance, Parameter	LSTM, SVM	propose a joint-learning cross-lingual approach to detect hate speech, encoding parallel source and target datasets via multilingual MUSE embeddings and integrating multilingual hate speech lexicon features (HurtLex) into LSTM networks	✓
[144]	2019	Instance, Parameter	BERT, mBERT	propose a multi-channel BERT (MC-BERT) model by translating source language to English and Chinese as parallel training or test data inputs and feeding into three versions of BERT (mBERT, English BERT and Chinese BERT)	✗
[41]	2019	Parameter	LSTM	build an LSTM-based model for the code-switched languages Hinglish by saving transferred weights across datasets for further training	✓
[42]	2019	Instance, Parameter	LSTM	experiment cross-lingual methods from source (English) or merged datasets to target (French) using multilingual distributional embeddings like Glove bilingual embeddings and self-mapped FastText embeddings	✗
[43]	2019	Instance	NB, RF, SVM	Experiment with the use of machine translation tools to translate test data to English and exploit different traditional models (such as SVM, naive Bayes, and random forest)	✗
[44]	2018	Instance, Parameter	CNN	create Hindi-English Offensive Tweet (HEOT) dataset, and experiment a transfer learning approach by training CNN on English data and part of translated Hinglish data, and then save transfer weights of CNN to re-train it on Hinglish data from HEOT	✗

Table A.2: Summary of cross-lingual techniques included in the automated identification of hate speech phenomena. "Ref" = "reference", and "Avail?" = "whether or not the codes and resources of the work are available".

B

SWSR Dataset Format

SWSR dataset consists of two files: "SexWeibo.csv" and "SexComment.csv", containing weibos (posts) and comments (replies) respectively. See more detailed description of features below:

B.1 SexWeibo.csv

- weibo_id: a string of weibo ID
- weibo_text: a string of weibo content
- keyword: contains sexism-related keyword(s) extracted from the weibo text
- user_gender: the gender of user
- user_location: the location of user
- user_follower: number of users who follow this user's account
- user_following: number of users whom this user follows
- weibo_like: number of like for the weibo
- weibo_comment: number of comment for the weibo
- weibo_repost: number of repost for the weibo
- weibo_date: the date and time when the weibo is posted

B.2 SexComment.csv

- weibo_id: the weibo id where the comment is collected
- comment_text: a string of the comment

- gender: the gender of commenter
- location: the location of commenter
- like: number of like for this comment
- date: the date and time when the comment is posted
- label: the comment is sexist(1) or non-sexist(0)
- category: categorise sexism into four classes – Stereotype based on Appearance(SA), Stereotype based on Cultural Background (SCB), MicroAggression (MA) and Sexual Offense (SO)
- target: the type of target who are attacked – Individual (I) or Generic (G)