applied sciences

Article

Detecting Offensive Language on Malay Social Media: A Zero-Shot, Cross-Language Transfer Approach Using Dual-Branch mBERT

Xingyi Guo *, Hamedi Mohd Adnan * and Muhammad Zaiamri Zainal Abidin

Department of Media and Communication Studies, University of Malaya, Kuala Lumpur 50603, Malaysia; zaiamrizainal@um.edu.my

* Correspondence: 22060166@siswa.um.edu.my (X.G.); hamedi@um.edu.my (H.M.A.)

Abstract: Social media serves as a platform for netizens to stay informed and express their opinions through the Internet. Currently, the social media discourse environment faces a significant security threat—offensive comments. A group of users posts comments that are provocative, discriminatory, and objectionable, intending to disrupt online discussions, provoke others, and incite intergroup conflict. These comments undermine citizens' legitimate rights, disrupt social order, and may even lead to real-world violent incidents. However, current automatic detection of offensive language primarily focuses on a few high-resource languages, leaving low-resource languages, such as Malay, with insufficient annotated corpora for effective detection. To address this, we propose a zero-shot, cross-language unsupervised offensive language detection (OLD) method using a dual-branch mBERT transfer approach. Firstly, using the multi-language BERT (mBERT) model as the foundational language model, the first network branch automatically extracts features from both source and target domain data. Subsequently, Sinkhorn distance is employed to measure the discrepancy between the source and target language feature representations. By estimating the Sinkhorn distance between the labeled source language (e.g., English) and the unlabeled target language (e.g., Malay) feature representations, the method minimizes the Sinkhorn distance adversarially to provide more stable gradients, thereby extracting effective domain-shared features. Finally, offensive pivot words from the source and target language training sets are identified. These pivot words are then removed from the training data in a second network branch, which employs the same architecture. This process constructs an auxiliary OLD task. By concealing offensive pivot words in the training data, the model reduces overfitting and enhances robustness to the target language. In the end-to-end framework training, the combination of cross-lingual shared features and independent features culminates in unsupervised detection of offensive speech in the target language. The experimental results demonstrate that employing cross-language model transfer learning can achieve unsupervised detection of offensive content in low-resource languages. The number of labeled samples in the source language is positively correlated with transfer performance, and a greater similarity between the source and target languages leads to better transfer effects. The proposed method achieves the best performance in OLD on the Malay dataset, achieving an F1 score of 80.7%. It accurately identifies features of offensive speech, such as sarcasm, mockery, and implicit expressions, and showcases strong generalization and excellent stability across different target languages.

Keywords: cross-language model; offensive language detection; mBERT; unsupervised methods; transfer learning

check for updates

Citation: Guo, X.; Adnan, H.M.; Abidin, M.Z.Z. Detecting Offensive Language on Malay Social Media: A Zero-Shot, Cross-Language Transfer Approach Using Dual-Branch mBERT. *Appl. Sci.* **2024**, *14*, 5777. https://doi.org/10.3390/app14135777

Academic Editors: Francisco De Arriba-Pérez, Silvia García-Méndez and Enrique Costa-Montenegro

Received: 6 June 2024
Revised: 27 June 2024
Accepted: 28 June 2024
Published: 2 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of mobile network technology and the emergence of different social media platforms, Internet penetration rates continue to rise. The Internet has permeated people's daily lives and work, offering significant convenience through various applications for travel, shopping, communication, and more [1]. In the virtual society constructed by the Internet, netizens not only act as information recipients but also drive the widespread dissemination of information. Among these platforms, social media has experienced rapid growth in recent years, serving as a space for netizens to stay informed about current events, express opinions, and share snippets of their lives. However, on current social media platforms, some netizens exhibit a narrow perspective, hold extreme views, and are easily swayed by emotionally charged news reports, leading to the expression of offensive comments [2,3]. Offensive language is a broad concept encompassing various terms such as abusive language, cyber-bullying, hate speech, irony, profanity, and more, depending on the data source or viewpoint definition [4].

In recent years, numerous measures have been introduced to address hate speech and offensive language. On 25 February 2019, United Nations Secretary-General Guterres unveiled the "United Nations Strategy and Plan of Action on Hate Speech" to tackle the escalating global issues of xenophobia, racism, violence against women, anti-Semitism, and anti-Muslim hatred [5]. From Rwanda to Bosnia and Cambodia, offensive and hate speech has evolved into heinous crimes, including ominous signs of ethnic cleansing [6]. Amidst the recent COVID-19 pandemic, some extremists have utilized Twitter to disseminate racist remarks. Swift moderation by Twitter promptly eradicated these messages, underscoring the platform's commitment to combatting racial speech [7]. In Malaysia, social media has also witnessed instances of violent incidents triggered by offensive language. In 2020, a 20-year-old girl, Thivya Nayagi Rajendran, from Penang, tragically took her own life, leaving a note citing cyberbullying as the cause. The girl and her Bangladeshi male colleague faced online harassment after uploading a dance video online. In 2023, Malaysian actress Lee Yuan Ling survived a suicide attempt by hanging and courageously shared her experience on social media, highlighting the horrors of online violence. That same year, a 27-year-old Chinese–Malaysian singer, Yuki Koh, fell victim to a fatal stabbing by an obsessive male fan. The assailant, in his 40s, had relentlessly pursued the victim on social media for years, demonstrating the profound impact of such behaviors not only on individuals but on society as a whole.

Currently, research on offensive language detection (OLD) primarily focuses on high-resource languages, such as English, due to abundant dataset resources, monolingual dictionaries, and pre-trained language models [8,9]. However, on social media platforms, diverse languages are often used for offensive language, including different national languages, ethnic languages, and regional dialects [10]. Research on OLD in low-resource languages faces significant challenges due to the lack of labeled training data instances. Early methods for detecting offensive comments mainly employed dictionaries and traditional machine learning (ML) tools. Dictionary-based approaches recorded malicious vocabulary for detection, while traditional ML tools relied on manual feature engineering to extract text features, construct ML models, and perform offensive comment detection [11]. However, these methods struggled to promptly maintain and update dictionaries for large volumes of user comments. Traditional ML methods required labor-intensive feature engineering, resulting in time-consuming efforts and models with room for improved generalization. In recent years, due to the superior practical results of deep learning (DL) models, OLD methods have predominantly shifted towards DL-based methods [12]. In the English context, ample high-quality annotated data facilitates effective DL model training. However, in the Malay language scenario, there is a lack of publicly available large-scale datasets for offensive comments, posing challenges for training DL models.

Cross-lingual transfer (CLT) technique seeks to address the aforementioned challenge by transferring knowledge across languages, enabling models to accomplish tasks in a low-resourced target language with annotated samples from another source language [13]. The scenario of interest is zero-shot CLT, where the target language lacks any annotated samples [14]. This constitutes the most common and challenging issue in CLT research, and it is the focus of this paper. Specifically, our attention is on enhancing the performance of OLD task in zero-shot CLT scenarios. Recent findings highlight the remarkable CLT capabilities of Multilingual BERT (mBERT), a model pre-trained on Wikipedia data in 104 languages. Fine-tuning mBERT using annotated samples from the source language has emerged as a mainstream approach for cross-lingual text label prediction [15]. Given the practical scenario where unlabeled samples in the Malay language are easily obtainable and contain relevant information about the target distribution, we propose that jointly utilizing annotated samples from the source language (English) and unlabeled samples from the target language (Malay) for fine-tuning mBERT is an effective strategy to enhance cross-lingual OLD performance.

Detecting offensive speech in low-resource languages faces two primary challenges. Firstly, the limited availability of resources hampers the effective semantic encoding of text in these languages. Secondly, it impedes the effective training on offensive features inherent to these languages. Currently, most approaches to detecting offensive speech in low-resource languages rely on cross-lingual, pre-trained models. The foremost advantage of such methods lies in their ability to leverage unsupervised cross-lingual pre-training, thereby enabling detection in low-resource languages. However, previous methods have often overlooked the linguistic disparities between the source and target domains during transfer learning, consequently undermining detection performance in the target domain.

Building on the analysis above, this paper proposes a dual-branch CLT framework based on mBERT, designed to facilitate zero-shot transfer learning from high-resource languages (e.g., English) to low-resource languages (e.g., Malay), thereby enhancing the detection of offensive speech in low-resource languages. The main contributions of this paper include:

(1) Utilizing the first network branch of the mBERT model to extract features from both source and target domain data. By adversarially minimizing the Sinkhorn distance, effective domain-shared features are obtained.
(2) Identifying offensive pivot words through an unsupervised model and removing these pivots from the training data of the second network branch, thereby constructing an auxiliary offensive speech detection task. This encourages the model to rely on contextual information, enhancing detection performance in low-resource languages.
(3) Constructing a Malay offensive speech detection dataset, analyzing the knowledge graph of various offensive pivots, and introducing statements with or without pivot words to assess the model's understanding of deep semantics and its applicability in real-world scenarios.

The structure of the remainder of this paper is as follows. Section 2 reviews related research on OLD filed, including traditional manual feature detection methods, machine learning approaches, and deep learning techniques. Section 3 provides a detailed description of the proposed dual-branch zero-shot cross-lingual OLD framework based on Sinkhorn distance and pivots extraction. Section 4 validates the effectiveness of the proposed method through extensive experiments on low-resource Malay text for offensive speech detection. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Related Works

In the early stages of social media development, platforms employed human moderators and encouraged users to report offensive content. Such manual approach was not only inefficient but also challenging to handle the vast number of online comments, exposing moderators to daily negative sentiments and potential psychological harm. Consequently, researchers sought to automate the OLD task. Traditional methods relied on manually designed grammar features and rule-based matching for identifying offensive comments. Razavi et al. [16] extracted features at different conceptual levels, enhancing multi-level classification accuracy with statistical models and rule-based pattern matching. Xu et al. [17] constructed a comment dataset from the English video platform YouTube. They proposed an automated sentence-level filtering method utilizing grammar relationships between words to create filtering rules for matching and removing offensive language.

Chen et al. [18] manually extracted lexical and syntactic features, combined with a designed syntax rule list to match and identify offensive content. Although these studies extensively analyzed rule patterns for offensive language, the definition of offensive statements remains ambiguous. Such methods heavily rely on rule lists, making them suitable for comments with specific keywords or distinct patterns. However, they face challenges in detecting more subtle forms of offensive language.

In order to better study the characteristics of offensive comments and enhance the accuracy of automated detection, researchers have constructed numerous large datasets from various social media scenarios for analysis. They have also begun utilizing ML-based methods for detection. Bassignana et al. [19] established the multilingual hate word lexicon, HurtLex, starting with a corpus of over 1000 harmful words in Italian, enriching the target language lexicon. Davidson et al. [20] gathered random tweets from Twitter, employing n-gram features, part-of-speech (PoS) features, and sentiment score features. These features were applied to various classifiers, including Linear Regression (LR), Naive Bayesian (NB), Decision Trees (DT), and linear Support Vector Machine (SVM). Experimental results demonstrated the superior performance of LR and linear SVM. Watanabe et al. [21] proposed an ML-based approach for detecting offensive language in Twitter based on writing patterns, unigrams, and sentimental features. The work refined various rules for extracting features, including rules for extracting sentiment features such as the total score of positive/negative words, the number of positive/negative slang words, and the number of positive/negative emoticons. Van Hee et al. [22] collected data from social media sites containing English and Dutch, refining label categories, including threats, curses, insults, defamation, and explicit content. SVM were used as classifiers in classification experiments, incorporating features such as character-based and word-based N-gram features, vocabulary features, and topic features. In OLD tasks, temporal information, contextual information, and sentiment information play crucial roles. However, ML-based methods might overlook sequential relationships in feature extraction, limiting their ability to fully utilize contextual information. In addition, such methods face challenges in extracting complex text features, and cannot handle the issue of data sparsity well [23].

In recent years, DL-based methods have started to be applied in the field of OLD. DL algorithms can automatically learn sample features and analyze the data accordingly. When employing DL-based methods in natural language processing (NLP), the text needs to be processed into suitable word representations as input for the network. The next step involves constructing an appropriate network to train and learn from the input data [24]. Badjatiya et al. [25] compared Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) models with traditional ML tools for detecting hate speech on Twitter. The results demonstrated the superior performance of DL models over ML models. Pavloupos et al. [26] conducted a detection analysis of user comments on news portals and Wikipedia, identifying those containing abusive language. Leveraging a Greek news portal, they built a dataset of 1.6 million user comments, employing a Recurrent Neural Network (RNN) with an attention mechanism for classification. The attention mechanism automatically highlights potentially offensive words in comments. Park et al. [27] combined CharCNN and WordCNN to create a hybrid CNN model for OLD. Considering the uniqueness of Twitter tweets, the model simultaneously learned features at the character and word levels. Sigurbergsson et al. [28] built a Danish language dataset, employing an LSTM-based method to detect hate speech, cyberbullying, and other offensive language. Mishra et al. [29] applied Graph Convolutional Networks (GCN) in OLD tasks. By introducing additional information such as user personal data and social networks to address the limited features of short texts, they input data as graph nodes into GCN for learning. Through graph convolutional networks, the model could directly capture user language behavior and group structure features. Experimental results indicated that the combination of LR and GCN yielded the best results.

Recent studies have demonstrated the effectiveness of pre-trained language models like BERT in downstream tasks such as OLD [30]. Zhu et al. [31] utilized fine-tuned BERT for downstream classification tasks to identify offensive tweets. Howard et al. [32] successfully implemented transfer learning in offensive language detection using BERT and the ULMFiT (Universal Language Model Fine-tuning for Text Classification) method. Due to superior performance, such approaches have become mainstream for addressing OLD tasks. In the 2019 OffensEval competition, 7 out of the top 10 teams used BERT, differing mainly in parameter settings and preprocessing steps. As mBERT exhibits great CLT capabilities. it has gradually become the foundation for recent cross-lingual text detection and classification tasks. Kudugunta et al. [33] demonstrated the extraction of relevant features from cross-lingual pre-trained models for downstream tasks like part-of-speech tagging and named entity recognition, providing language-specific knowledge-based information. Wu et al. [34] employed a meta-learning approach to achieve improved cross-lingual transfer with M-BERT, particularly in named entity recognition. However, this method requires re-finetuning the model for each test sample during the inference stage, resulting in substantial computational costs. Keung et al. [35] used language adversarial training to encourage mBERT to generate more language-independent features. Zhang et al. [36] considered CLT tasks as domain adaptation problems, and utilized an unsupervised domain adaptation method based on adversarial training to reduce domain differences. Kumar et al. [37] achieved significant results in German and Hindi OLD tasks from social media by fine-tuning the mBERT pre-trained model. Libovický et al. [38] confirmed that context-based mBERT captures similarities between languages, clustering them by language, and cross-lingual fine-tuning does not disrupt this property. In other words, mBERT encodes part of the language information based on its position in the embedding space, concentrating the encoding of each language to achieve a degree of cross-lingual capability. Ayo et al. [39] proposed a method employing SVM and mBERT to construct a cross-lingual model for detecting offensive and misogynistic language on Twitter. Kapil et al. [40] enhanced mBERT with transfer learning strategy, transferring knowledge from a resource-rich language to low-resource languages, effectively improving the accuracy of OLD in low-resource languages. However, these methods overlook the differences between the source and target languages during the transfer learning process, impacting detection performance in the target domain. This oversight leads to negative transfer effects, which compromise detection performance in the target domain.

Negative transfer refers to the phenomenon in transfer learning where the transfer of knowledge or models from a source domain to a target domain not only fails to improve performance but may actually degrade it [41]. Significant linguistic differences exist between source and target languages, encompassing grammar, vocabulary usage habits, and more. These differences can hinder effective model transfer in zero-shot scenarios [42]. Variations in language usage backgrounds and cultural contexts across different languages can affect the model's understanding of text meaning, thereby impacting the effectiveness of transfer learning [43]. Disparities in data distribution between source and target languages, including differences in data volume, types, and quality, can also lead to poor model performance in the target language [44]. To mitigate the risks of negative transfer, this paper leverages the multilingual pre-training capability of mBERT. This enables the model to learn universal language representations. By minimizing the Sinkhorn distance, a stable feature alignment mechanism is established between source and target languages. This helps ensure the model effectively maintains consistency in the feature space during transfer, reducing negative transfer due to language differences. Furthermore, handling and masking pivot words during training enhances the model's robustness to the target language. Additionally, a joint domain-shared feature training strategy allows the model to simultaneously consider and utilize information from both source and target languages. This effectively addresses potential negative transfer issues in zero-shot CLT learning, enhancing the model's generalization capability and stability.

In the realm of cross-lingual OLD, ML models such as LR and SVM offer high interpretability, crucial for understanding how models make predictions and which features influence these predictions most significantly [45]. LR classifies by learning the weights of features, directly interpreting each feature's weight as its contribution to the final classification. This transparency makes LR intuitive and straightforward in terms of interpretability, as its linear decision boundary visibly shows which combinations of feature values lead to specific classification decisions. SVM's decisions hinge on support vectors—samples on the decision boundary—determining not only its location but also aiding in understanding the model's certainty and accuracy in delineating between different classes. When using kernel functions like the Radial Basis Function (RBF) kernel, SVM maps data into high-dimensional space to find the optimal decision hyperplane, where kernel choice affects model interpretability. In contrast, DL models are often dubbed "black box" models due to their complex internal structures and parameters, complicating direct explanations of decision processes and feature selection principles [46]. The interpretability of DL models tends to be more indirect, often relying on a deeper understanding of their internal mechanisms. For instance, models like mBERT, a pretrained language model, operate with intricate internal mechanisms learned via self-attention on vast text datasets. Consequently, the meanings of its features may not be as intuitively clear as those derived from traditional linear models like LR and SVM. Moreover, transfer learning emphasizes sharing knowledge across different tasks or domains, where interpretability hinges on effectively selecting and transferring knowledge from the source domain and adapting it to the target domain. Feature alignment and transfer rules play pivotal roles in this process. For example, our approach utilizes Sinkhorn distance minimization to enhance feature alignment between source and target languages, mitigating the risk of negative transfer caused by language disparities and thereby improving model generalization. To summarize, LR and SVM excel in interpretability due to their intuitive feature weight explanations and visualizable decision boundaries. In contrast, models like mBERT and transfer learning methods, leveraging large-scale pretrained models and complex transfer strategies, may offer interpretability that is more indirect or reliant on a deeper understanding of internal mechanisms. Therefore, a balance between interpretability and model performance must be struck when selecting the most suitable method to address specific task requirements.

3. Proposed Framework

In the realm of cross-language text processing, a prevalent challenge lies in the linguistic disparities between source and target languages. If common offensive features can be identified across different languages, and unique language-specific attributes are thoroughly considered in offensive determinations, it can play a pivotal role in cross-language OLD tasks. We propose a dual-branch zero-shot, cross-language OLD model based on mBERT combined with Sinkhorn distance and pivot extraction.

For a given set of monolingual offensive language instances, the first network branch trains on mBERT model with source language (e.g., English) offensive language samples to create a monolingual offensive language detector, then it is transferred to the given low-resource language (e.g., Malay). By estimating Sinkhorn distance between source and target language feature representations, adversarially minimizing the Sinkhorn distance provides more stable gradients, extracting more effective cross-domain shared features.

Subsequently, an auxiliary prediction task is constructed. In the input training data of source (labeled) and target (unlabeled) language for the second network branch, offensive pivot words are concealed to compel the model to rely more on contextual information for determining text offensiveness. This process yields domain-independent features, aiding the model in better understanding the context of the instance, thereby enhancing its capability to handle complex language structures and implicit information. Through joint training of the entire network, domain-shared features and domain-specific features are amalgamated to fulfill cross-language OLD tasks. Figure 1 illustrates the framework flowchart of the proposed method.

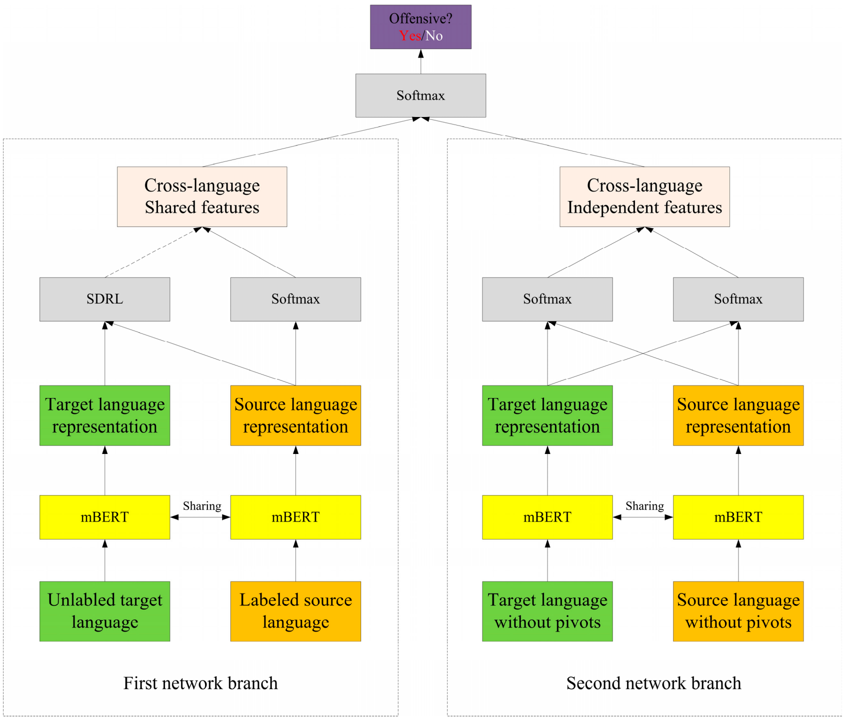

Figure 1. Illustration of the proposed dual-branch, zero-shot CLT framework for OLD tasks. In the first network branch on the left side, the labeled data from the source domain and the unlabeled data from the target domain are jointly used as inputs for feature extraction. mBERT transfer learning incorporates SDRL (Sinkhorn distance representation layer) to extract cross-language shared features. On the right side, the second network branch removes offensive pivot words from training data in both source and target languages, obtaining language-specific independent features. Ultimately, through end-to-end framework joint training of the two network branches, the OLD performance is enhanced for the unlabeled target language.

3.1. Problem Definition

In cross-language OLD tasks, let D_s and D_t be the training data of source and target languages, respectively. D_s contains labeled data $X_s^l = \{x_s^i, y_s^i\}_{i=1}^{N_s^l}$, where N_s^l denotes the quantity of annotated data in the source language, with y_s representing the labels for offensive detection in the source domain. D_t comprises unlabeled data $X_t = \{x_t^j\}_{j=1}^{N_t}$, where N_t is the quantity of unlabeled data in the target domain. The proposed method aims to leverage labeled samples from the source language to train a classifier for detecting offensive text in unlabeled samples X_t from the target domain.

3.2. BERT

The BERT model obtains semantically richer word embeddings through pre-training and fine-tuning, overcoming the limitations of traditional word embeddings' polysemy issue [47]. Fine-tuning allows the model to be applied to specific downstream tasks such as OLD, enhancing its generalization ability [48].

For pre-training, a base model is constructed by stacking the encoder part of Transformer models, as illustrated in Figure 2. Joint training of Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks captures semantic vector representations at both word and sentence levels. The MLM achieves the true bidirectional language modeling [49]. In the transfer to downstream tasks, BERT draws inspiration from OpenAI's Generative Pre-training (GPT) model and designs more versatile input and output layers, surpassing the generality of GPT [50].

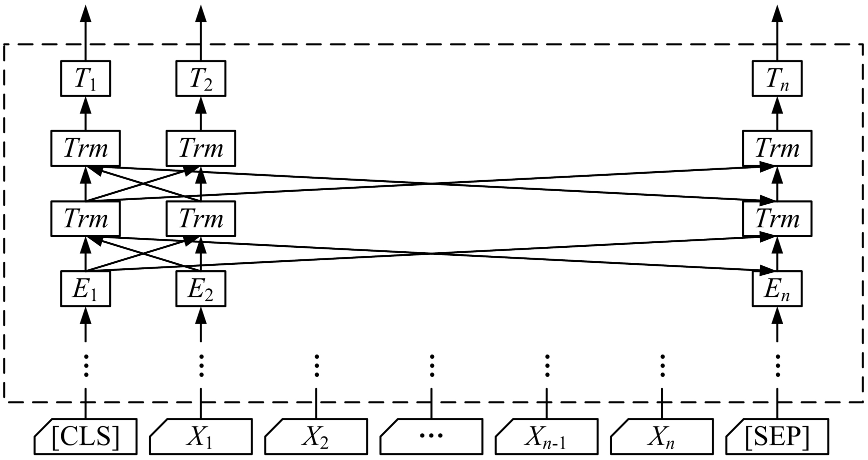

Figure 2. Model structure of BERT.

In BERT, text information is first transformed into word embeddings E, and then input into the Transformer structure, consisting of self-attention mechanisms and feedforward neural networks (FFN). The BERT pre-trained model undergoes extensive pre-training on large corpora, enabling it to convert semantically related words or characters into one-dimensional word vectors with close distances in the feature vector space. This retains more valuable information and emotional tendencies, facilitating the differentiation of offensive comments in subsequent tasks. In typical scenarios, the degree of association between words with negative sentiments and personal pronouns in a sentence largely determines its malicious intent. For instance, in the sentence written in Malay, "Kubur sudah sedia untukmu, kenapa kamu tidak masuk?" (The grave is ready for you, why do you not go in?), there are no explicit vulgar words. However, the word "kubur" (grave) carries strong negative connotations and is closely related to the personal pronoun "kamu" (you), demonstrating an offensive attitude. Therefore, inputting the transformed word vectors into the Transformer structure and calculating the closeness of their relationship aids in determining whether the statement is offensive.

The obtained word vectors E are used as input, passing through multiple layers of Transformer feature extractors, yielding feature information (T_1, T_2, ..., T_n). During the self-attention mechanism, an initial random matrix X is generated. The word embeddings multiplied by X produce three new vectors: Query Q, Key, and Value, denoted as Q, K, V. The corresponding Q, K, and V vectors for multiple word embeddings form matrices W^Q, W^K, and W^V. The use of matrices W simplifies the computation process. The Q of each input word embedding is computed with the K of other word embeddings, and both word vectors have an initial V. The operation result of Q and K generates a weight, merging the importance of the corresponding relationships between two words into the initial V, thereby preserving contextual semantic information. For example, to determine the relationship between "kubur" (grave) and "kamu" (you) in the sentence, the dot product of Q_{kubur} and K_{kamu} is calculated, resulting in a score S:

$$S = Q \cdot K \tag{1}$$

By extension, a set of scores S = $\{S_1$, S_2, ..., $S_n\}$ can be obtained, representing the closeness of relationships between specific words (e.g., "kubur") and all other words in the text. Dividing this score set by a constant and performing softmax operation, followed by multiplication with V, yields the feature T that encapsulate the relationships between words. The constant is typically the square root of the arithmetic dimensions of the initial random matrix:

$$T = soft\max(\frac{S}{\sqrt{d_k}})V \tag{2}$$

where d_k represents the initial random matrix dimensions, V is the Value vector, and T is the output vector. The output vector encapsulates implicit features indicating the current word's association with other words. In subsequent training, these features assist in discerning offensive comments. The aforementioned computation calculates the relationship between one word and others. Multiple such computations are performed simultaneously to compute pairwise relationships between each pair of words. This calculation is crucial for recognizing offensive comments. After training with a substantial amount of offensive text, the model becomes adept at identifying offensive comments that do not necessarily contain explicit vulgar words.

Supervised learning methods typically require ample annotated data, which is available for English across various tasks. However, for languages with limited resources, acquiring abundant labeled data for supervised training poses challenges. Therefore, the transition from a well-trained high-resource language to a low-resource one, without the need for annotated data, becomes a new challenge. Unsupervised CLT methods, being domain-independent and not reliant on annotated data, are well-suited for languages with large-scale open and unstructured data.

3.3. mBERT

mBERT, a multilingual version of BERT, is trained on Wikipedia texts from 104 languages, allowing it to handle multilingual text within the same representation space, and a lot of studies have highlighted mBERT's excellent CLT performance [51]. Hence, this paper adopts mBERT as the base model.

A linear classification layer <W, b> can be directly added to the mBERT model f_θ to build the model ψ, $\psi = \{f_\theta, W, b\}$. Given a labeled dataset $X_s^l = \{x_s^i, y_s^i\}_{i=1}^{N_s^l}$ in the source language, the model is trained using the dataset X_s^l, resulting in the base model.

$$\psi_0 = \underset{\psi}{\operatorname{argmin}} \sum_{x_s^i \in X_s^l} CE(p(x_s^i, \psi), y_s^i) \tag{3}$$

where CE is the cross-entropy function. $p(x_s^i, \psi)$ represents the label distribution of model output samples. With mBERT, being a multilingual pre-trained model, ψ_0 can be directly applied to OLD task in the target language, and it generally offers some cross-language generalization. However, the direct transfer faces challenges and drawbacks:

(1) Language differences: Significant variations in syntax, vocabulary, and expressions exist across languages. Without labeled data in the target language, the model may struggle to adapt to specific language structures and cultural nuances, leading to performance degradation.

(2) Domain disparities: Even within the same language family, domain differences can persist. Applying a model trained in one domain directly to another may fail to capture specific features of the target domain, impacting performance.

(3) Task specifics: OLD is a task with unique challenges. Expression of offensive language can vary widely across languages and cultures. Models trained directly in the source language might not effectively capture offensive language features in the target language.

(4) Label bias: Offensive language in the source and target languages may have different label distributions. Without target language labels, the model may encounter label bias issues, resulting in unstable performance in the target language.

(5) Limitations of transfer learning: Despite mBERT's inherent transfer learning capabilities, without labeled data in the target language, the model's transfer ability may be restricted, making it challenging to fully adapt to the specific requirements of the target language.

3.4. Reducing Cross-Domain Discrepancy Based on Sinkhorn Distance

Similarity measurement is crucial for evaluating the degree of similarity between feature vectors. For feature vectors with ample representational capabilities, the choice of a similarity measurement algorithm directly impacts classification accuracy [52]. In metric learning, commonly used algorithms include Euclidean distance and cosine distance. Euclidean distance emphasizes absolute differences in numerical values between vectors, making it sensitive to outliers [53]. Cosine distance, on the other hand, focuses on the relative differences in the direction of vectors, remaining insensitive to absolute values [54]. However, cross-domain text features are prone to noise, category-independent feature points, and feature distribution that fails to meet the assumption of independent identical distribution. In such cases, using Euclidean or Cosine distance is inadequate. To address this, we employ the Box-Cox transformation to redistribute feature vectors extracted by mBERT into Gaussian or quasi-Gaussian distributions [55]. We then use the Sinkhorn distance as the similarity measurement for feature vectors [56].

Sinkhorn distance, proposed by Cuturi M, based on optimal transport, differs from traditional similarity measurement algorithms like Euclidean and cosine distances [57]. It calculates the work required to move from one vector to another, offering a unique perspective compared to direct physical distance calculations. The Sinkhorn distance can be seen as an optimal mapping from the source domain distribution to the target domain distribution, aiding in the understanding of inter-domain differences while offering greater stability compared to traditional metrics like Maximum Mean Discrepancy (MMD) that may be susceptible to local optima. It effectively handles high-dimensional, nonlinear data distribution disparities. This capability is particularly crucial in natural language processing, where text data often exhibit intricate semantic and syntactic structures, challenging for conventional distance measures to capture. Moreover, the Sinkhorn distance serves not only as a measure between domains but also supports domain adaptation learning. Minimizing the Sinkhorn distance between source and target domains facilitates learning more generalized feature representations, thereby enhancing model performance on the target domain. Importantly, grounded in optimal transport theory, the Sinkhorn distance provides robust theoretical underpinnings and mathematical guarantees. Under suitable conditions, it ensures convergence to genuine distribution disparities, a critical factor in several practical applications.

The computation process of Sinkhorn distance is illustrated in Figure 3, where $P(x)$ and $Q(x)$ represent two vectors, and the arrows depict the movement from $P(x)$ to $Q(x)$.

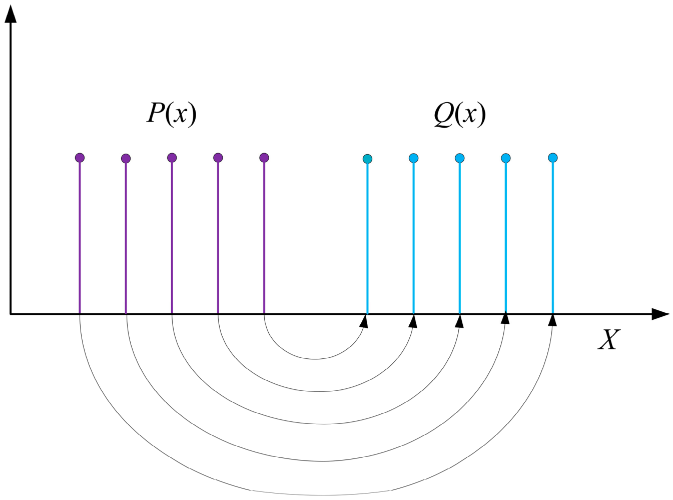

Figure 3. Illustration of Sinkhorn distance.

The aforementioned optimal transport process is defined as follows:

$$d(a,b) = \sum_{s\in U(a,b)} < S, L > = \sum_{i,j} S_{i,j} L_{i,j} \tag{4}$$

where S represents a coupling matrix. a and b are probability distributions of vectors P and Q, respectively, ensuring that the sums of elements in a and b are both 1. $U(a,b) = \left\{S \in R_{+}^{n\times m}\right\}$ denotes that U is an $n \times m$ matrix with rows summing to a and columns summing to b. L_{ij} signifies the cost of moving from point i to point j.

Since S is a probability distribution matrix, while Sinkhorn distance holds a natural advantage when the vector distributions are similar or approximately similar, complexities arise in computations, particularly with higher-dimensional vectors and extensive data. To address this, an entropy regularization term is added to the equation:

$$d(a,b) = \sum_{i,j} S_{i,j} L_{i,j} + \beta H(S) \tag{5}$$

where, β is the entropy regularization parameter, and $H(S)$ is the entropy function of S:

$$H(S) = -\sum_{i,j} S_{i,j} \log(S_{i,j}) \tag{6}$$

Approximating Sinkhorn distance typically involves maximizing a loss function L_{wd} concerning the parameter θ_w. This loss function is designed as a dual problem that minimizes Sinkhorn distance. Specifically, for the representation distributions P and Q of the source and target domains, along with their corresponding data X and Y, maximizing the Sinkhorn distance loss function with respect to the parameter θ_w can be expressed as:

$$L_{wd}(\theta_w) = \max_{u,v}\left(\sum_i u_i + \sum_j u_j - \varepsilon \sum_{i,j} u_i e^{-\varepsilon c_{i,j}} v_j\right) \tag{7}$$

where u and v are scaling coefficients. ϵ is the regularization parameter, and $c_{i,\,j}$ represents the cost or distance between samples X_i and Y_j. During the process of maximizing L_{wd}, the scaling coefficients u and v are iteratively updated to approximate the actual Sinkhorn distance.

Implementing a gradient penalty on the parameters is an effective approach to enforcing Lipschitz constraints [58]. This is particularly useful in DL-based models, controlling the model's Lipschitz constant through gradient clipping or regularization, enhancing stability, and mitigating issues like gradient explosion or vanishing during training [59].

In the context of Sinkhorn distance, enforcing Lipschitz constraints is typically achieved by adding a gradient penalty term to the loss function. The gradient penalty term can be in the form of parameter norms or other forms to ensure that the model's gradients do not become excessively large during training. To ensure the model's Lipschitz continuity, we choose the gradient penalty term $L_{grad}(\theta_w)$ as the L2 norm of the model parameters:

$$L_{grad}(\theta_w) = \frac{1}{2}|\theta_w|_2^2 \tag{8}$$

Due to the Sinkhorn distance being nearly continuously differentiable, the initial step involves training SDRL. After reaching optimal parameters for SDRL, these parameters are fixed, and the estimation of Sinkhorn distance is minimized. This process enables the mBERT network to learn domain-shared feature representations, accomplished by solving an optimization problem involving both maximization and minimization:

$$L_{adv} = \min_{\theta_{mBERT}} \max_{\theta_w} \{L_{wd} - \lambda L_{grad}\} \tag{9}$$

where θ_{mBERT} represents the parameters of the feature extraction layer. λ is the penalty coefficient, set to 0 during the minimization process, ensuring the gradient penalty does not guide the learning process. The optimization goal is to learn domain-shared feature representations by minimizing the Sinkhorn distance loss function L_{wd}. During training, the minimization of L_{wd} is achieved by adjusting the parameters θ_{mBERT} of the feature extractor network through solving a maximization-minimization problem. The estimation parameters θ_w for Sinkhorn distance are determined by maximizing L_{wd}. The overall optimization problem involves two components:

(1) Minimizing Sinkhorn distance: Adjusting the parameters θ_{mBERT} of the feature extractor network to reduce the loss of Sinkhorn distance.

(2) Maximizing gradient penalty: Tuning the estimated parameters θ_w for Sinkhorn distance to increase the gradient penalty term λL_{grad}. The ultimate optimization goal is to strike a balance between these two components, achieving a model with both domain-shared feature representation and compliance with Lipschitz constraints. This type of optimization problem is commonly referred to as adversarial training, where one network (m*BERT*) aims to minimize the loss, while the other network (*SDRL*) seeks to maximize it. This adversarial training process aids in learning more discriminative feature representations while satisfying predefined constraints.

Upon completing the training phase focused on reducing domain differences using Sinkhorn distance, the text representation d is mapped to label y through a softmax classifier for discerning offensive language. The loss L_{mBERT} of the *mBERT* classifier employs the cross-entropy loss function. Combining this with the loss of the SDRL classifier yields the objective function for the first network branch:

$$L_1 = \min_{\theta_{mBERT},\theta_{SDRL}} \{L_{mBERT} + \beta \max_{\theta_w}[L_{wd} - \lambda L_{grad}]\} \tag{10}$$

where β acts as the coefficient controlling the balance between offensive detection and learning domain-shared features.

3.5. Extracting Domain-Independent Features Based on Pivot Words

When there are fewer shared features between the source and target languages, the performance of cross-language OLD tends to decline. Effectively utilizing domain-specific features enhances detection capabilities. Typically, domain-specific pivot words in the dataset co-occur with domain-shared pivot words that exhibit similar offensive tendencies. To extract domain-specific features, it is essential to construct an auxiliary task closely related to the original task.

Firstly, the PERL method proposed in the paper [60] is employed to select domain-shared feature words, known as pivots. Words that meet the criteria of being adjectives, adverbs, or verbs with high attention weights are chosen as pivots for attack judgment. Subsequently, pivots are concealed in the training data of both the source and target languages, and a special UNK tag is used to replace all the pivots in the original data.

In BERT, the UNK (Unknown) label is employed to denote Out-Of-Vocabulary (OOV) terms, referring to words not encountered during the model's training, absent from the pre-trained vocabulary. Using the UNK label aids the model in better generalizing across different languages. Given BERT's capacity to grasp contextual information during pre-training, it is anticipated that the model can infer the semantic details of UNK-labeled words through context, thereby enhancing performance in tasks like cross-language text analysis. This approach also mitigates the model's reliance on pivot words for learning task labels, facilitating improved adaptation to new languages.

The CLT training of mBERT, concealing pivot words and training a model for detecting offensive text, offers several advantages:

(1) Enhanced Generalization: By concealing pivot words, the trained model avoids over-reliance on specific vocabulary, thereby improving its ability to generalize to a broader range of offensive text and increasing overall applicability.

(2) Reduced Overfitting: Over-dependence on pivot words during training may lead to overfitting, where the model performs well on training data but poorly on unseen data.

Hiding these crucial words diminishes overfitting tendencies, enhancing robustness to new data.

(3) Improved Contextual Understanding: Hiding pivot words compels the model to rely more on contextual information to discern text offensiveness. This fosters a better understanding of textual context, improving the model's capability to handle complex language structures and implicit information.

(4) Enhanced Model Robustness: Concealing pivot words during training increases the likelihood of the model learning more universal and general features of offensive text, bolstering its robustness across different contexts.

We introduce an auxiliary label, y_{off}, for each transformed data instance. If the original data instance contains at least one offensive pivot, label y_{off} as 1; otherwise, label it as 0. The transformed data serves as the input to the feature extraction layer of the second network branch to generate feature representations, predicting the auxiliary label y_{off}. The loss function is denoted as L_{off}. By training on the prediction task of auxiliary labels, the second network branch can extract domain-specific features coinciding with domain-shared features.

3.6. Incorporation of Both Domain-Shared and Domain-Specific Features

To better leverage domain-shared and domain-specific features, the two branches of the proposed framework undergo joint training. Source domain data and transformed data serve as inputs to the feature extraction layer, yielding feature representations d_s and $d_{s'}$. Concatenating both for OLD classification training with the loss function L_{joint}. The overall objective loss function combines the loss terms and regularization from both network branches:

$$L = L_{joint}(d_s \oplus d_{s'}) + L_{wd} + L_{off} + \vartheta L_{reg} \tag{11}$$

where ϑL_{reg} is an L2 regularization term applied to the parameters of mBERT, SDRL, and auxiliary label prediction to prevent overfitting. ϑ is a parameter used to balance the regularization term with other terms.

4. Experiment and Analysis

4.1. Data Collection

To assess model performance, a comprehensive analysis was conducted on the proposed method in executing zero-shot unsupervised detection of Malay offensive language. We collected a dataset of Malay offensive language from Twitter to support the training and evaluation of OLD models. The dataset was meticulously constructed to encompass tweets containing specific Malay offensive keywords, alongside contrasting tweets devoid of such keywords. This step aimed at creating a representative dataset covering various facets of offensive language.

Initially, a specific set of Malay offensive keywords was identified, addressing diverse aspects of offensive discourse, including but not limited to personal attacks, racial discrimination, or malicious statements. Examples include terms like "Bodoh" (Stupid/Foolish), "Mampus kau" (You are dead), "Lelaki jamban" (Men are unpleasant as a toilet), "Kafir" (Arabic term often used in Islamic contexts to refer to a non-believer or someone who rejects or disbelieves in Islam), and "Bapok" (Used to insult or demean transgender individuals or effeminate gay men). Leveraging Twitter's API and its keyword filtering functionality, we collected tweets containing these offensive keywords. To construct the comparative dataset, we utilized the Twitter API to gather tweets that did not include offensive keywords.

In the realm of cyberbullying, personal attacks emerge as its primary manifestation and a focal point in OLD tasks. Detecting discriminatory language proves challenging, relying on models to autonomously discover underlying patterns within the corpus. Subsequently, the collected tweets undergo data cleansing to eliminate duplicates, invalid information, and irrelevant content. Text preprocessing, encompassing tasks like tokenization and stop-word removal, enhances the efficacy of subsequent model training. The final dataset comprises approximately 625 offensive tweets and about 2200 non-offensive tweets.

It is crucial to note that the presence of offensive words in a tweet does not inherently classify it as offensive. Manual scrutiny of each tweet, considering the evolving connotations of specific terms within niche online cultures, is imperative. Comments with explicit vulgarities are directly categorized as offensive. As for texts with strong malicious intent but lacking explicit offensive terms, their classification often relies on subjective human perception. As of now, there are no universally quantifiable rules, and the majority sentiment determines the standard. Three individuals participated in the dataset annotation process, adopting a one-person labeling and two-person verification approach to ensure consistency in judgment criteria. In cases of differing opinions, a voting mechanism was employed to determine the classification outcome.

For the source language of model training, this paper utilizes the English annotated dataset from the 2019 OffensEval shared task 6, known as EN-OLID [61]. Previous efforts predominantly focused on detecting specific types of malicious content, neglecting to explore commonalities across different types of offenses (e.g., personal insults termed as cyberbullying and insults targeting groups termed as hate speech). This dataset takes a comprehensive approach, proposing the OLID (Offensive Language Identification Dataset) annotations. It employs a hierarchical modeling approach, discerning whether comments are malicious, identifying the type of malicious comments (non-targeted or targeted), and pinpointing the target of malicious comments (individuals or groups). These hierarchical annotations capture commonalities across diverse offensive text data, rendering it applicable to various OLD tasks.

To comprehensively assess the effectiveness and generality of the proposed method in cross-language OLD task, experiments also employed Danish and Arabic datasets released in the 2020 OffensEval shared task 12 [62]. Table 1 provides details on the sample distribution in the experimental datasets, while Table 2 presents examples of offensive texts in different languages.

Table 1. Data distribution of the experiment dataset.

Language	**Training Set**		**Testing Set**		**Sum**
	Offensive	**Normal**	**Offensive**	**Normal**	
English	4400	8840	240	620	14,100
Malay	500	1760	125	440	2825
Danish	344	2320	40	256	2960
Arabic	1395	5660	155	629	7839

Table 2. Offensive instances for different languages.

Language	**Offensive Instances**
English	You are really dumb, I cannot believe you made such a stupid comment. Seriously?
Malay	Kau benar-benar bodoh, tak boleh percaya kau buat komen bodoh macam tu. Seriuslah?
Danish	Du er virkelig dum, kan ikke tro, du sagde det.
Arabic	أنت غبي حقًا، لا أستطيع أن أصدق أنك قلت ذلك.

4.2. Parameter Configuration

This experiment was conducted using the PyTorch framework, with training performed on an NVIDIA GeForce RTX 2080 Ti GPU. Specific parameter settings are outlined as follows. For mBERT model, the word vector dimension is set at 768. This dimension is sufficient to capture rich semantic information and grammatical features, making it suitable for processing multilingual textual data. The mBERT pre-trained model utilizes a vocabulary of 104 languages, totaling 120,000 words. Through data analysis, it is observed that the majority of samples have lengths within 140 words; hence, the maximum sentence length was set to 140. The Softmax layer has hidden units corresponding to the 2 label (offensive/normal) categories. During model training, mBERT's word embedding layer remains fixed. Due to its pre-training on large-scale data, mBERT has learned rich semantic information in its word embedding layer. Therefore, in specific tasks such as cross-lingual OLD studied in this paper, the word embedding layer is typically frozen to prevent overfitting and expedite model convergence. The batch size is set to 32, with a total of 10 training epochs. The choice of batch size typically involves balancing computational resources and training efficiency. A smaller batch size, such as 32, can accelerate convergence and enhance the model's generalization ability. The number of epochs (10) signifies how many times the model will see the entire training dataset, which is usually sufficient for the model to converge to an appropriate state. The optimizer used is AdamW [63], with a weight decay of 0.01 and a learning rate of 3×10^{-5} to control the size of gradient updates. A linear warm-up of learning rate is applied for the first 10% of training steps, gradually increasing the learning rate during the initial stages to stabilize the model training process. Dropout rate is set to 0.1, this helps reduce the risk of overfitting the model while maintaining its generalization ability.

4.3. Evaluation Metrics

To evaluate the cyberbullying detection performance of the model, precision, recall, and F_1 score are used as the primary evaluation metrics:

$$Pre = \frac{TP}{TP + FP} \tag{12}$$

$$Rec = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{14}$$

where TP denotes instances correctly identified as offensive, while TN signifies correctly identified normal instances. Conversely, FN refers to offensive instances mistakenly classified as normal, and FP represents normal instances incorrectly classified as offensive. The F_1 score, a holistic assessment metric, integrates precision and recall for a comprehensive evaluation.

Furthermore, to introduce error analysis metrics in OLD, we use the Brier Score (BS) to measure the accuracy of probability predictions. The calculation of BS for cross-lingual OLD is:

$$BS = \frac{1}{N}\sum\nolimits_{i=1}^{N} (p_i - y_i)^2 \tag{15}$$

where N is the total number of samples. p_i represents the model's predicted probability that sample i is offensive. y_i is the ground truth label for sample i, where y_i = 1 indicates the text is offensive, and y_i = 0 indicates it is non-offensive. The BS measures the mean squared error between the model's predicted probabilities and the actual labels, evaluating the performance of the cross-lingual OLD model. BS ranges from 0 to 1, with lower values indicating more accurate probability predictions; 0 represents perfect predictions, and 1 represents the worst predictions.

Log Loss (LL) is used to measure the discrepancy between the predicted probabilities of a model and the actual labels. Although the BS focuses more on the accuracy of predicted probabilities, LL places greater emphasis on the divergence between predicted probability distributions and the true labels. LL is calculated as:

$$LL = -\frac{1}{N}\sum\nolimits_{i=1}^{N} [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{16}$$

LL calculates the logarithmic loss between the predicted probability distribution of the model and the actual labels. A lower value indicates that the model's predictions are closer to the true labels, with 0 representing perfect prediction accuracy.

4.4. Comparison Methods

In the experimental validation of the proposed method for OLD in Malay, several comparison models are set up for comparison:

1. LR (Logistic Regression): Utilizing TF-IDF for text feature extraction, followed by LR detector for constructing an offensive language classifier. LR is a linear model, simple and interpretable. In specific parameter settings, feature extraction employs TF-IDF with a maximum of 5000 features and an n-gram range of (1, 2). The minimum document frequency (min_df) is set to 0.05, and the maximum document frequency (max_df) to 0.95. For LR, the penalty type is "l2" with a regularization strength of 1.0, using the solver "liblinear" with a maximum of 100 iterations [64].
2. SVM: Extracting features from ext using TF-IDF and feeding them into a SVM detector to learn text features. SVM classifies text by finding the optimal hyperplane in feature space. SVM classifies text by finding the optimal hyperplane in the feature space. The regularization parameter for SVM is set to 1.0, with a radial basis function kernel (rbf) and a gamma coefficient set to "scale" [65].
3. mBERT (Unsupervised): Training mBERT model on labeled source language dataset and leveraging mBERT's multilingual transfer capabilities for direct application in OLD task in the target language. However, limitations exist due to language differences, task specificity, label bias, and constraints of transfer learning.
4. MMSE [66]: Pretraining a multilingual neural translation model and transferring through the encoder for model transfer.
5. MDD-VAT [36]: Enhancing mBERT using unsupervised domain adaptation with target language unlabeled samples through virtual adversarial training.
6. mBERT (Supervised): Training mBERT model on labeled target language dataset for benchmark performance reference. In supervised settings, mBERT undergoes end-to-end training using labeled data for OLD tasks in the target language.

4.5. Results and Discussions

Table 3 presents the performance results on the Malay language test dataset using different models. Except for mBERT (Supervised), all other methods utilize Malay unlabeled samples as training data. Results show that LR and SVM methods, which directly extract features from target language unlabeled samples without employing transfer learning techniques, perform the worst. While ML-based classifiers have the advantages of simplicity, intuitiveness, and ease of implementation, they struggle to capture complex semantics and context relationships, lacking support for transfer learning and adaptability to language and task variations.

Table 3. OLD performance with different models (/%).

Models	Precision/%	Recall/%	F1/%	BS	LL
LR	52.1	40.2	33.6	0.49	0.82
SVM	62.1	50.9	44.9	0.44	0.78
mBERT (Unsupervised)	60.2	52.9	48.7	0.40	0.66
MMSE [66]	73.5	71.4	69.4	0.28	0.47
MDD-VAT [36]	76.7	74.4	72.3	0.25	0.42
Proposed method	81.5	80.5	80.7	0.14	0.23
mBERT (Supervised)	80.6	79.4	79.1	0.16	0.25

mBERT (Unsupervised), with its multilingual transfer capabilities and ability to train without target language labels, captures shared features to some extent. However, it fails to address issues such as language differences, domain disparities, and label bias.

The proposed method outperforms other unsupervised approaches in all metrics. This is attributed to using mBERT as the base language model in a dual-branch CLT framework. By minimizing Sinkhorn distance adversarially, the model extracts more effective domain-shared features, enhancing knowledge transfer in cross-language tasks. Additionally, constructing auxiliary OLD tasks aids the model in learning more universal and general offensive text features. The strategy of hiding offensive pivots in the training data forces the model to rely more on contextual information to determine text offensiveness. This approach helps the model better adapt to text features in the target language. The proposed method's superior performance stems from these considerations in addressing challenges like language differences and complex semantics.

In contrast to the supervised approach mBERT (Supervised), which excels in adapting to OLD tasks in the target language by utilizing labeled data, it requires a substantial amount of labeled data for effective training. This can be particularly challenging for low-resource languages where obtaining sufficient labeled data is a hurdle. Therefore, mBERT (Supervised) may struggle to adapt to the specific contexts and expressions of offensive language in low-resource languages when faced with insufficient labeled data. In contrast, our proposed method employs a transfer learning framework, performing zero-shot transfer from a high-resource language to a low-resource language, enabling unsupervised offensive text detection. This reduces the reliance on extensive labeled data. By hiding offensive pivots in the second branch of the model and constructing auxiliary OLD tasks, our method compels the model to depend more on contextual information to assess text offensiveness. This aids in learning more universal and general offensive text features. Results indicate that our method's performance in Malay OLD tasks is slightly superior to mBERT (Supervised).

In order to further assess the proposed model's zero-shot transfer learning capability for OLD tasks in different languages, we conducted experiments using English as the annotated source language and Danish and Arabic from the 2020 OffensEval shared task 12 dataset as unlabeled target languages. The results, illustrated in Figure 4, demonstrate the model's performance in unsupervised OLD across different languages. Compared to the results in Malay, it is evident that despite having the fewest unlabeled training samples in the target language and the most in Arabic, the proposed method performs best on the Danish dataset and least effectively on Arabic. This discrepancy arises due to the significant impact of semantic similarity on transfer learning efficacy. The closer the target language is to the source language, the better the knowledge transfer. Danish and English share some commonalities as they both belong to the Germanic language family, exhibiting certain similarities in grammar, vocabulary, and language structure. On the other hand, Arabic and Malay belong to different language families, showing greater differences from English in terms of grammar and structure. Additionally, Arabic is a right-to-left written language, while English is left-to-right, further increasing the dissimilarity between them. Malaysia, being a culturally diverse country with Malay, Chinese, Indian, and other ethnic groups, has a unique multicultural influence shaped by neighboring countries, history, and religion. In contrast, Arabic countries predominantly follow Islamic culture. The cultural disparities between Arabic-speaking countries and English-speaking ones may contribute to the decrease in model transfer learning performance.

The proposed method incorporates pivot extraction and hiding strategies in the training data, utilizing a second network branch for auxiliary label prediction tasks to assist the model in capturing language-specific features of offensive speech. Offensive pivots play a crucial role in OLD tasks. These pivots are key terms in the text that often carry malicious, insulting, or hostile semantics. By detecting and analyzing these pivot words, the model can more accurately identify and classify offensive speech. Therefore, the presence of pivot words can serve as one of the bases for text classification by the model. Figure 5 illustrates the distribution of offensive pivots in the Malay language dataset. Table 4 provides details on pivot extraction in the dataset along with some examples.

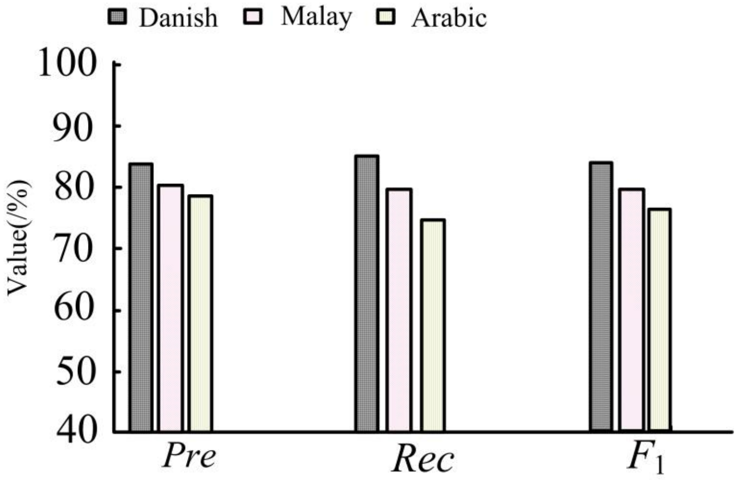

Figure 4. Zero-shot OLD results of the proposed method across different languages.

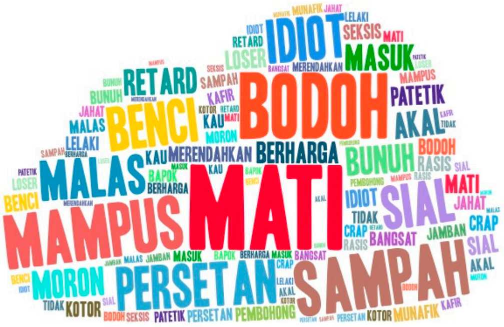

Figure 5. Visualization pivots distribution.

Table 4. Instances from the collected dataset and pivot extraction examples.

Sample Number	Sentence	Pivots	Offensive?
S1	Kau ni tak guna, macam bangsat. (You are useless, like scum.)	bangsat	yes
S2	Laman web anda adalah satu penipuan. Jangan percayakan penipu ini. (Your website is a scam. Do not trust this scammer.)	penipuan, penipu	yes
S3	Dia buat perkara yang agak bodoh tetapi sebenarnya dia punya hati yang baik. (He did something rather foolish, but actually he has a good heart.)	Bodoh	no
S4	Kebijaksanaanmu nampaknya hanya mencukupi untuk melakukan perkara-perkara mudah, seperti bernafas. (Your intelligence seems to be sufficient only for doing simple things, like breathing.)	/	yes

The presence of aggressive pivot words may indicate that the entire text is more likely to be offensive. For instance, in the examples S1 and S2 from Table 4, the model can determine both sentences as offensive using pivots like “bangsat”, “penipuan”, and “penipu”. However, the appearance of aggressive pivot words might introduce noise, leading the model to misinterpret some sentences. Some texts contain aggressive pivot words, but the overall context is not offensive, resulting in false positives. Some offensive statements may lack common aggressive pivot words, causing the model to miss them. Additionally, some texts may not contain aggressive pivot words but still possess offensive content. Overreliance on pivot words may cause the model to overlook such cases. Moreover, in transfer learning, certain pivot words that are offensive in the source language may not carry the same meaning in the target language, leading to misjudgments. For example, in the instance S3 from Table 4: "Sejujurnya, saya baru mencuba resipi masakan baru, tetapi rasa yang dihasilkan agak bodoh. Mungkin saya perlu menambah lebih rempah pada masa akan datang!" ("To be honest, I just tried a new recipe, but the taste it produced was a bit bland. Maybe I need to add more spices in the future!"), the use of the word "Bodoh" is to express a bland taste rather than for offensive purposes. In Malay, "Bodoh" is sometimes used to describe something's underperformance but usually without insult. Similarly, some sentences may lack offensive words but still be aggressive using sarcasm or mockery, as seen in comment S4 from Table 4: "Intelek kamu nampaknya hanya mencukupi untuk melakukan perkara yang mudah, seperti bernafas". ("Your intelligence seems to be sufficient only for doing simple things, like breathing"). This statement is highly offensive as it includes derogatory remarks about the other person's intelligence. Aggression is not solely about specific words but also encompasses the overall tone and intent of the language. When constructing a model for detecting offensive text, one must consider such indirect forms of aggression. In our experiments, the comparative methods failed to correctly identify the offensiveness of the examples S3 and S4. In contrast, our proposed method successfully determined the offensiveness of these sentences through the joint training strategy of hiding pivot words and the domain-specific feature extraction capability in the auxiliary OLD task.

Finally, we analyze the impact of varying training resources on the transfer effect. We take different quantities of labeled training data from the source language (English), including 1000, 2000, 5000, and all (13,240) samples, and examine the performance of the proposed method on the Malay dataset. Figure 6 presents the F1 results, maintaining the training and testing data volume of non-tabled Malay samples consistent with the previous experiments. The results reveal that even with just 1000 source language training samples, the proposed method's performance surpasses the earlier experiment's direct transfer of the mBERT (Unsupervised) model. This demonstrates the method's high sample utilization efficiency. A larger quantity of labeled source language data implies a richer knowledge base for the model. This knowledge can be effectively transferred to the target language in transfer learning, providing a better starting point and helping the model adapt to the target language's context and expressions. Abundant source language data contributes to establishing more robust and universal feature representations, crucial for OLD models, as offensive language might exhibit similar features across different languages. In summary, the proposed model leveraging annotated samples from high-resource languages can achieve more robust and accurate OLD in the target language, providing substantial support for handling low-resource language data.

In order to further evaluate the effectiveness of each component module proposed in this study, we conducted ablation experiments. Figure 7 presents performance results on the Malay test dataset, where all models were trained using annotated English texts and unlabeled Malay texts, with Malay texts as the test data. Model 1 solely employed the unsupervised mBERT model, utilizing its multilingual transfer capabilities directly for detecting offensive text in the target language. Model 2 used only the first network branch of the proposed method (left branch in Figure 1), replacing the SDRL layer with a Gradient Reversal Layer (GRL), a common technique in domain adaptation aimed at reducing feature differences between different domains (such as source and target domains) to enhance model generalization in the target domain. The basic idea behind GRL is to force the network to learn feature representations that do not contain domain-specific information by reversing the direction of gradients. Model 3 employed only the first network branch of the proposed method. Model 4 represents the complete model of the proposed method, which constructs auxiliary tasks based on pivot words to extract domain-independent features and integrates cross-lingual shared features and independent features in an end-to-end training framework to achieve unsupervised detection of offensive speech in the target language. From the results, Model 1 performs the worst, indicating that the basic mBERT model struggles with effective cross-domain knowledge transfer. Model 2, by reversing gradient directions, forces the network to learn domain-invariant feature representations, enabling the model to learn more generalized feature representations in domain adaptation and thereby reducing domain discrepancies to some extent, resulting in improved performance. Model 3 achieves suboptimal performance through adversarial minimization of Sinkhorn distance, suggesting that methods based on Sinkhorn distance are more effective than those based on GRL for reducing domain discrepancies. Finally, the complete proposed model (Model 4) achieves the best performance, validating that the proposed method effectively aids in learning more universal and general offensive text features by constructing auxiliary OLD tasks, demonstrating its effectiveness in zero-shot cross-lingual OLD tasks.

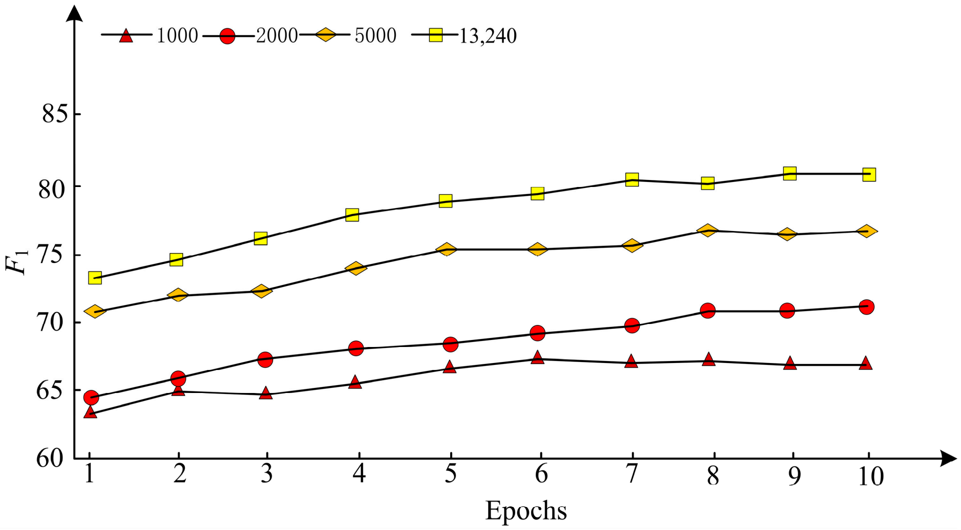

Figure 6. F_1 results for OLD task on the Malay dataset with varying amounts of source language training data.

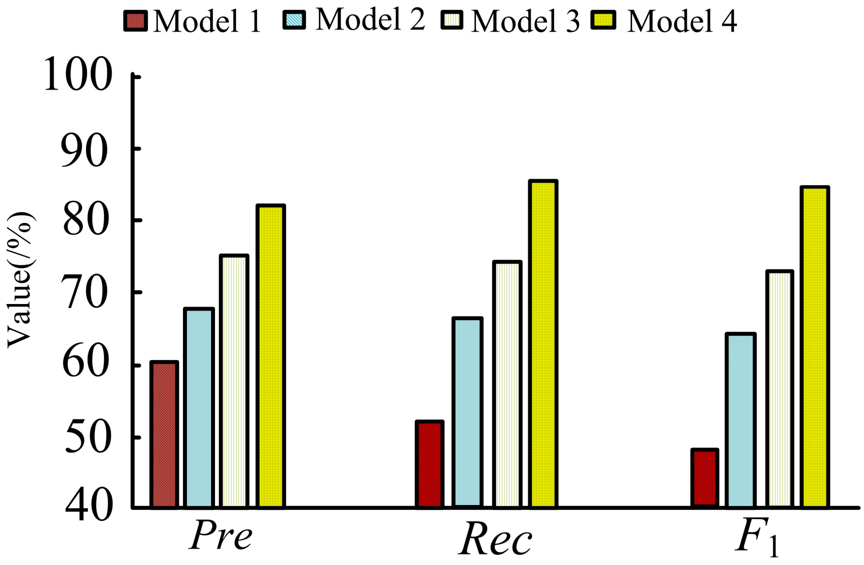

Figure 7. Results of the ablation experiments.

5. Conclusions

In addressing the shortcomings of existing methods in feature extraction for low-resource languages, such as Malay, in OLD tasks, this paper introduces a dual-branch zero-shot cross-lingual model based on mBERT. Effectively leveraging the Sinkhorn distance, the model acquires cross-lingual shared features with strong discriminatory power for detecting offensive content. Augmented with an auxiliary label prediction task, unique domain-specific features are extracted. The combination of these features facilitates robust training and achieves superior performance in offensive text detection compared to existing methods. The success of this approach can be attributed to the comprehensive utilization of the multilingual pre-trained model mBERT, efficient feature extraction strategies, adept adversarial training, and a thoughtful approach to handling offensive pivot words and dataset construction. Collectively, these factors contribute to the outstanding performance of this method in unsupervised CLT for offensive text detection.

Moreover, the proposed method may also face certain limitations. The interpretability of the adopted mBERT model is notably poor, restricting the model's credibility and explainability in specific scenarios. Despite employing unsupervised learning methods, the construction of datasets for OLD still relies on labeled data from high-resource languages such as English. The process of constructing and handling offensive pivot words could potentially render the model sensitive to specific attributes within the dataset. In our future research endeavors, we aim to explore the development and application of visualization tools such as LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to aid in elucidating the model's decision-making processes and feature importance. Additionally, we intend to investigate active learning strategies to automatically select the most informative samples for annotation, thereby reducing dependency on extensive annotated data from the source language. Furthermore, by introducing more variations and noise from real-world scenarios into our datasets, we seek to mitigate the model's reliance on specific data attributes.

In summary, the proposed approach demonstrates significant advancements in addressing OLD challenges in low-resource languages, providing valuable insights for the application of CLT learning in safeguarding security and order in social media environments. As further improvements in model performance and generalization are pursued, the potential impact of this method in practical applications is expected to grow. Future research endeavors will explore the application of this approach to various tasks, including machine translation, text generation, and other areas within the field of NLP.

Author Contributions: Conceptualization, X.G., H.M.A. and M.Z.Z.A.; Methodology, X.G., H.M.A. and M.Z.Z.A.; Software, X.G.; Validation, X.G., H.M.A. and M.Z.Z.A.; Investigation, X.G.; Resources, X.G. and M.Z.Z.A.; Writing—original draft, X.G.; Writing—review & editing, H.M.A. and M.Z.Z.A.; Supervision, H.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data are part of an ongoing study.

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

References

1. Aichner, T.; Grünfelder, M.; Maurer, O.; Jegeni, D. Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 215–222. [CrossRef] [PubMed]
2. Meel, P.; Vishwakarma, D.K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Syst. Appl.* **2020**, *153*, 112986. [CrossRef]
3. Barrutia, L.; Vega-Gutiérrez, J.; Santamarina-Albertos, A. Benefits, drawbacks, and challenges of social media use in derma-tology: A systematic review. *J. Dermatol. Treat.* **2022**, *33*, 2738–2757. [CrossRef] [PubMed]

4. Risch, J.; Ruff, R.; Krestel, R. Offensive language detection explained. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 11–16 May 2020; pp. 137–143.
5. Baider, F. Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech. *Politics Gov.* **2023**, *11*, 249–260. [CrossRef]
6. Marques, T. The expression of hate in hate speech. *J. Appl. Philos.* **2023**, *40*, 769–787. [CrossRef]
7. Awan, I.; Carter, P.; Sutch, H.; Lally, H. Online extremism and Islamophobic language and sentiment when discussing the COVID-19 pandemic and misinformation on Twitter. *Ethn. Racial Stud.* **2023**, *46*, 1407–1436. [CrossRef]
8. Jahan, M.S.; Oussalah, M. A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing* **2023**, *546*, 126232. [CrossRef]
9. Chinivar, S.; Roopa, M.S.; Arunalatha, J.S.; Venugopal, K.R. Online offensive behaviour in social media: Detection approaches, compre-hensive review and future directions. *Entertain. Comput.* **2023**, *45*, 100544. [CrossRef]
10. Mahmud, T.; Ptaszynski, M.; Eronen, J.; Masui, F. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Inf. Process. Manag.* **2023**, *60*, 103454. [CrossRef]
11. Akhter, M.P.; Zheng, J.; Naqvi, I.R.; AbdelMajeed, M.; Zia, T. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimed. Syst.* **2022**, *28*, 1925–1940. [CrossRef]
12. Tontodimamma, A.; Nissi, E.; Sarra, A.; Fontanella, L. Thirty years of research into hate speech: Topics of interest and their evolu-tion. *Scientometrics* **2021**, *126*, 157–179. [CrossRef]
13. Pikuliak, M.; Šimko, M.; Bieliková, M. Cross-lingual learning for text processing: A survey. *Expert Syst. Appl.* **2021**, *165*, 113765. [CrossRef]
14. Martínez-García, A.; Badia, T.; Barnes, J. Evaluating morphological typology in zero-shot cross-lingual trans-fer. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1, Long Papers), Virtual, 1–6 August 2021; pp. 3136–3153.
15. Muller, B.; Elazar, Y.; Sagot, B.; Seddah, D. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. *arXiv* **2021**, arXiv:2101.11109.
16. Razavi, A.H.; Inkpen, D.; Uritsky, S.; Matwin, S. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence, Proceedings of the 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, ON, Canada, 31 May–2 June 2010*; Proceedings 23; Springer: Berlin/Heidelberg, Germany, 2010; pp. 16–27.
17. Xu, Z.; Zhu, S. Filtering offensive language in online communities using grammatical relations. In Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, Washington, DC, USA, 13–14 July 2010; pp. 1–10.
18. Chen, Y.; Zhou, Y.; Zhu, S.; Xu, H. Detecting offensive language in social media to protect adolescent online safety. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; IEEE: New York, NY, USA, 2012; pp. 71–80.
19. Bassignana, E.; Basile, V.; Patti, V. Hurtlex: A multilingual lexicon of words to hurt. In Proceedings of the CEUR Workshop Proceedings, CEUR-WS, Turin, Italy, 18 September 2018; Volume 2253, pp. 1–6.
20. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11, pp. 512–515.
21. Watanabe, H.; Bouazizi, M.; Ohtsuki, T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive ex-pressions and perform hate speech detection. *IEEE Access* **2018**, *6*, 13825–13835. [CrossRef]
22. Van Hee, C.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Automatic detection of cyberbullying in social media text. *PLoS ONE* **2018**, *13*, e0203794. [CrossRef] [PubMed]
23. Mahmud, T.; Das, S.; Ptaszynski, M.; Hossain, M.S.; Andersson, K.; Barua, K. Reason based machine learning approach to detect bangla abusive social media comments. In Proceedings of the International Conference on Intelligent Computing & Optimization, Hua Hin, Thailand, 27–28 October 2022; pp. 489–498.
24. Alrashidi, B.; Jamal, A.; Khan, I.; Alkhathlan, A. A review on abusive content automatic detection: Approaches, challenges and opportunities. *PeerJ Comput. Sci.* **2022**, *8*, e1142. [CrossRef] [PubMed]
25. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep learning for hate speech detection in tweets. In Proceedings of the 26th Inter-national Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 759–760.
26. Pavlopoulos, J.; Malakasiotis, P.; Androutsopoulos, I. Deeper attention to abusive user content moderation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1125–1135.
27. Park, J.H.; Fung, P. One-step and two-step classification for abusive language detection on twitter. *arXiv* **2017**, arXiv:1706.01206.
28. Sigurbergsson, G.I.; Derczynski, L. Offensive language and hate speech detection for Danish. *arXiv* **2019**, arXiv:1908.04531.
29. Mishra, P.; Del Tredici, M.; Yannakoudakis, H.; Shutova, E. Abusive language detection with graph convolutional networks. *arXiv* **2019**, arXiv:1904.04073.
30. Chakkarwar, V.; Tamane, S.; Thombre, A. A Review on BERT and Its Implementation in Various NLP Tasks. In *International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)*; Atlantis Press: Dordrecht, The Netherlands, 2023; pp. 112–121.

31. Zhu, J.; Tian, Z.; Kübler, S. UM-IU@ LING at SemEval-2019 task 6, Identifying offensive tweets using BERT and SVMs. *arXiv* **2019**, arXiv:1904.03450.
32. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1, Long Papers), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018.
33. Kudugunta, S.R.; Bapna, A.; Caswell, I.; Arivazhagan, N.; Firat, O. Investigating multilingual NMT representations at scale. *arXiv* **2019**, arXiv:1909.02197.
34. Wu, Q.; Lin, Z.; Wang, G.; Chen, H.; Karlsson, B.F.; Huang, B.; Lin, C.Y. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 9274–9281.
35. Keung, P.; Lu, Y.; Bhardwaj, V. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. *arXiv* **2019**, arXiv:1909.00153.
36. Zhang, D.; Nallapati, R.; Zhu, H.; Nan, F.; dos Santos,, C.N.; McKeown, K.; Xiang, B. Margin-aware unsupervised domain adaptation for cross-lingual text labeling. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Virtual, 16–20 November 2020; pp. 3527–3536.
37. Kumar, A.; Saumya, S.; Singh, J.P. NITP-AI-NLP@ HASOC-FIRE2020, Fine Tuned BERT for the Hate Speech and Offensive Content Identification from Social Media. In Proceedings of the FIRE (Working Notes), Hyderabad, India, 16–20 December 2020; pp. 266–273.
38. Libovický, J.; Rosa, R.; Fraser, A. How language-neutral is multilingual BERT? *arXiv* **2019**, arXiv:1911.03310.
39. Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A. Hate speech detection in Twitter using hybrid embeddings and improved cuck-oo search-based neural networks. *Int. J. Intell. Comput. Cybern.* **2020**, *13*, 485–525. [CrossRef]
40. Kapil, P.; Ekbal, A. A deep neural network based multi-task learning approach to hate speech detection. *Knowl.-Based Syst.* **2020**, *210*, 106458. [CrossRef]
41. Yu, Y.; Huang, J.; Liu, S.; Zhu, J.; Liang, S. Cross target attributes and sample types quantitative analysis modeling of near-infrared spectroscopy based on instance transfer learning. *Measurement* **2021**, *177*, 109340. [CrossRef]
42. Meftah, S.; Semmar, N.; Tamaazousti, Y.; Essafi, H.; Sadat, F. On the hidden negative transfer in sequential transfer learning for domain adaptation from news to tweets. In Proceedings of the Second Workshop on Domain Adaptation for NLP, Kiev, Ukraine, 19–20 April 2021; pp. 140–145.
43. Alqahtani, Y.; Al-Twairesh, N.; Alsanad, A. A comparative study of effective domain adaptation approaches for arabic sentiment classification. *Appl. Sci.* **2023**, *13*, 1387. [CrossRef]
44. Kanclerz, K.; Miłkowski, P.; Kocoń, J. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Comput. Sci.* **2020**, *176*, 128–137. [CrossRef]
45. Hasib, K.M.; Rahman, F.; Hasnat, R.; Alam, M.G.R. A machine learning and explainable ai approach for predicting secondary school student performance. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 26–29 January 2022; IEEE: New York, NY, USA, 2022; pp. 0399–0405.
46. Buhrmester, V.; Münch, D.; Arens, M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 966–989. [CrossRef]
47. Garí Soler, A.; Apidianaki, M. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 825–844. [CrossRef]
48. Merendi, F.; Dell'Orletta, F.; Venturi, G. On the Nature of BERT: Correlating Fine-Tuning and Linguistic Competence. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 3109–3119.
49. Shah Jahan, M.; Khan, H.U.; Akbar, S.; Umar Farooq, M.; Gul, S.; Amjad, A. Bidirectional Language Modeling: A Systematic Literature Review. *Sci. Program.* **2021**, *2021*, 1–15. [CrossRef]
50. Yang, B.; Luo, X.; Sun, K.; Luo, M.Y. Recent progress on text summarisation based on bert and gpt. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Guangzhou, China, 16–18 August 2023; Springer Nature: Cham, Switzerland, 2023; pp. 225–241.
51. Mabokela, K.R.; Celik, T.; Raborife, M. Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access* **2022**, *11*, 15996–16020. [CrossRef]
52. Ge, L.; Parhi, K.K. Classification using hyperdimensional computing: A review. *IEEE Circuits Syst. Mag.* **2020**, *20*, 30–47. [CrossRef]
53. Dokmanic, I.; Parhizkar, R.; Ranieri, J.; Vetterli, M. Euclidean distance matrices: Essential theory, algorithms, and applications. *IEEE Signal Process. Mag.* **2015**, *32*, 12–30. [CrossRef]
54. Usino, W.; Prabuwono, A.S.; Allehaibi KH, S.; Bramantoro, A.; Hasniaty, A.; Amaldi, W. Document similarity detection using k-means and cosine distance. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 165–170. [CrossRef]
55. Osborne, J. Improving your data transformations: Applying the Box-Cox transformation. *Pract. Assess. Res. Eval.* **2019**, *15*, 12.
56. Bhardwaj, R.; Vaidya, T.; Poria, S. KNOT: Knowledge Distillation using Optimal Transport for Solving NLP Tasks. *arXiv* **2021**, arXiv:2110.02432.
57. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2292–2300.

58. Genevay, A.; Cuturi, M.; Peyré, G.; Bach, F. Stochastic optimization for large-scale optimal transport. *Adv. Neural-Form. Process. Syst.* **2016**, *29*.
59. Lyu, J.; Zhang, S.; Qi, Y.; Xin, J. Autoshufflenet: Learning permutation matrices via an exact lipschitz continuous penalty in deep convolutional neural networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 608–616.
60. Ben-David, E.; Rabinovitz, C.; Reichart, R. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embeding models. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 504–521. [CrossRef]
61. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Semeval-2019 task 6, Identifying and categorizing offensive language in social media (offenseval). *arXiv* **2019**, arXiv:1903.08983.
62. Zampieri, M.; Nakov, P.; Rosenthal, S.; Atanasova, P.; Karadzhov, G.; Mubarak, H.; Derczynski, L.; Pitenis, Z.; Çöltekin, Ç. SemEval-2020 task 12, Multilingual offensive language identification in social media (OffensEval 2020). *arXiv* **2020**, arXiv:2006.07235.
63. Yao, Z.; Gholami, A.; Shen, S.; Mustafa, M.; Keutzer, K.; Mahoney, M. Adahessian: An adaptive second order optimizer for machine learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 9 February 2021; Volume 35, pp. 10665–10673.
64. Gnanavel, S.; Duraimurugan, N.; Jaeyalakshmi, M.; Rohith, M.; Rohith, B.; Sabarish, S. A live suspicious comments detection using TF-IDF and logistic regression. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 4578–4586.
65. Gopi, A.P.; Jyothi, R.N.S.; Narayana, V.L.; Sandeep, K.S. Classification of tweets data based on polarity using improved RBF kernel of SVM. *Int. J. Inf. Technol.* **2023**, *15*, 965–980. [CrossRef]
66. Artetxe, M.; Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [CrossRef]

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