Cross-lingual Capsule Network for Hate Speech Detection in Social Media

Aiqi Jiang
Queen Mary University of London
London, UK
a.jiang@qmul.ac.uk

Arkaitz Zubiaga
Queen Mary University of London
London, UK
a.zubiaga@qmul.ac.uk

ABSTRACT

Most hate speech detection research focuses on a single language, generally English, which limits their generalisability to other languages. In this paper we investigate the cross-lingual hate speech detection task, tackling the problem by adapting the hate speech resources from one language to another. We propose a cross-lingual capsule network learning model coupled with extra domain-specific lexical semantics for hate speech (CCNL-Ex). Our model achieves state-of-the-art performance on benchmark datasets from AMI@Evalita2018 and AMI@Ibereval2018 involving three languages: English, Spanish and Italian, outperforming state-of-the-art baselines on all six language pairs.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Data analytics**; • **Computing methodologies** → **Natural language processing**; • **Social and professional topics** → *Women*.

KEYWORDS

Cross-lingual Learning; Capsule Network; Hate Speech Detection; Social Media

1 INTRODUCTION

Anonymity and lack of moderation provide benefits for social media, alongside negative effects such as production of harmful and hateful contents [18, 32, 39]. The necessity to tackle hate speech has attracted the attention of the scientific community, industry and government to come up with automated solutions. Most existing research focuses on English hate speech, due to its advantageous position over other languages in terms of available resources. This in turn leads to a dearth of research in other languages [18]. Despite the recent trend of increasingly investigating hate speech

detection in other languages [36], most of the work is limited to a single language. As far as we know, a few documented efforts have been made on cross-lingual hate speech detection in terms of multilingual word embeddings [1, 27, 28], multilingual multi-aspect hate speech analysis [26], and state-of-the-art cross-lingual contextual embeddings like multilingual BERT [14] and XLM-RoBERTa [19, 29]. To further research in broadening the generalisability and suitability of models across languages for hate speech detection [39], we propose a Cross-lingual Capsule Network Learning model with Extra lexical semantics specifically for hate speech (CCNL-Ex), whose two-parallel framework enriches inputs information in both source and target languages. The model can be applied to new languages lacking annotated training data [7, 28] and is able to capture spatial positional relationships between words to improve the generalisability of capsules. It can also be exploited to broaden the detection capacity on linguistically diverse genres such as social media. Our key contributions include: (1) we introduce the first approach to cross-lingual hate speech detection that incorporates capsule networks; (2) we integrate hate-related lexicons into pre-trained word embeddings to investigate their potential to further boost performance; (3) our model yields state-of-the-art performance for all six language pairs under study compared with ten baselines; (4) we perform a comparative study looking into the impact of each layer on our model.

2 RELATED WORK

Cross-lingual Hate Speech Detection. With the prevalence of online social media, a range of NLP approaches, especially transformer-based techniques [35], have been employed to identify online hate speech [1, 18, 39] or focusing on detecting specific types of hate, such as racism [13, 37], sexism [27, 38], and cyberbullying [5, 30], but limited to a single language, generally in English. While more mature NLP fields such as sentiment analysis have accumulated substantial efforts by employing cross-lingual learning techniques, work on hate speech detection in cross-lingual scenario has not been explored as much. Basile and Rubagotti [3] use SVM with n-grams to tackle English and Italian in a cross-lingual setting in Evalita 2018 and achieve the 15th/2nd position for English/Italian. Pamungkas and Patti [28] propose a joint-learning cross-lingual model with multilingual HurtLex [4] and MUSE embeddings [24], which outperforms other models using monolingual embeddings [28]. Several multilingual multi-aspect approaches are conducted for hate speech [26] and cross-lingual contextual word embeddings are applied in offensive language identification from English to other languages [19, 29]. Due to the scarcity of cross-lingual resources in this field, some studies tend to generate parallel corpora directly leveraging machine translation resources such as Google Translate [9, 27, 28, 41], which has proved the effectiveness of the
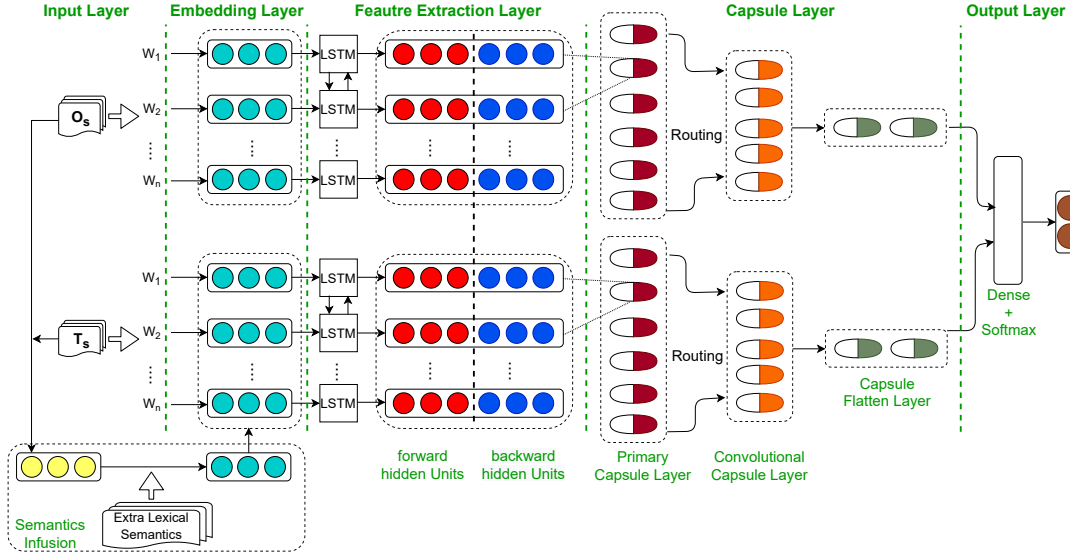
Figure 1: The architecture of CCNL-Ex.

approach. In our study, we consider Pamungkas and Patti [28]'s cross-lingual joint model as a baseline which also uses machine translation as part of their pipeline.

Capsule Networks. Capsule Network is a clustering-like method proposed by Sabour et al. [31]. They replace scalar-output feature detectors of CNNs with vector-output capsules to learn spatial relationships of entities via dynamic routing, improving representations against CNNs. Hinton et al. [22] then propose a new EM-based iterative routing, which shows potential in image analysis [31] and is soon applied to NLP research [40]. Its use on hate speech detection is however limited [33, 34]. Srivastava et al. [34] put forward a capsule-based architecture for aggressive language classification, and further incorporate multi-dimensional capsules for the same task [33]. Capsule network has not been considered in cross-lingual hate speech detection so far. Hence, we contribute to gaps in both lines of research bringing together cross-lingual hate speech detection and capsule network.

3 OUR PROPOSED METHOD: CCNL-EX

3.1 Model Architecture

Inspired by Capsule Networks built by Sabour et al. [31], we propose a cross-lingual capsule network learning model with multilingual word embeddings integrated lexical semantics (CCNL-Ex). It is composed of six layers (see Figure 1):

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

Embedding Layer. The input data is the sequence of texts and each text consists of a series of words. The input representation is a weight matrix $X \in \mathbb{R}^{e \times V}$ for e-dimensional vector of words and vocabulary size of V, fine-tuned by absorbing extra hate-related lexical semantic information (see Section 3.2).

Feature Extraction Layer. In each aligned network, we use a Bidirectional Long Short Term Memory (BiLSTM) network [21] as the feature extractor to get contextual relationships from local features. The output of BiLSTM is $h_t = [h_t^f, h_t^b] \in \mathbb{R}^{(2 \times k)}$, combined by forward feature h_t^f and backward feature h_t^b with k units.

Capsule Layer. The capsule layer consists of a primary capsule layer and a convolutional capsule layer. The primary capsule layer extracts instantiation parameters to represent spatial position relationships between features, like local order of words and their semantics [40]. Suppose $W \in \mathbb{R}^{(2 \times k) \times d}$ is a shared matrix, where d is the dimensionality of capsules. For the hidden feature h_t, we create each capsule $p_i \in \mathbb{R}^d$:

$$p_i = g(W^T h_t + b) \tag{1}$$

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

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

The convolutional capsule layer is connected to capsules in the primary capsule layer. Some primitive routing algorithms, like max pooling in CNNs, only capture features to show whether it exists in a certain position or not, missing more spatial relationships [40]. The connection weight is learnt by a dynamic routing, which can reduce the loss to make the capsule network more formative and effective, and attach less significance of unrelated or useless content, such as stop words [34]. The process of dynamic routing between the primary capsule u_i and the convolutional capsule v_j is as below:

$$c_{j|i} = softmax(b_{j|i}) \tag{3}$$

$$v_j = g(\sum_i c_{j|i} u_{j|i}) \tag{4}$$

$$b_{j|i} = b_{j|i} + u_{j|i} \cdot v_j \tag{5}$$

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

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

3.2 Lexical Semantic Knowledge Infusion

We fine-tune pre-trained word embeddings by infusing domain-specific lexical semantic knowledge, aiming to obtain domain-aware word representations and enhance model capacity of identifying hate-related content. More specifically, we firstly retrieve five most relevant semantic words from SenticNet [6] for each lexical word, and utilise Fasttext embedding model [20] to generate the five most similar words for each out-of-vocabulary (OOV) word. Then we apply the similarity learning method proposed by Faruqui et al. [15] to integrate lexicon-derived semantic information into pre-trained word embeddings by minimising distances between a word and its semantically related words.

4 EXPERIMENTS

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

4.1 Datasets

We use gender-based hate speech datasets from the Automatic Misogyny Identification (AMI) tasks held at the Evalita 2018[1] and IberEval 2018[2] evaluation campaigns. These datasets provided by AMI@Evalita and AMI@IberEval are extracted from the Twitter platform, and constructed under the same annotation scheme for binary labels: misogynistic and non-misogynistic. AMI@Evalita datasets present texts in English and Italian [16], while the AMI@Ib-erEval ones are in Spanish and English [17]. We utilise English data only from AMI@Evalita to make data size balanced among three languages, as well as for consistency with previous research [28] to enable direct comparison. In view of the well-divided training and test sets for each language, we further randomly select 20% of the training set as the validation set for model fine-tuning process, and finally utilise the whole training set to evaluate model capacity on test set. More details of datasets can be seen in Table 1. We create parallel corpora for separate datasets by directly using Google Translate[3] to translate all data between source and target languages.

4.2 Multilingual Lexicons

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

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

Table 1: Distribution of train, validation and test sets, misogynistic text rate (MTR) in source training and test sets, data sources for three languages.

Language	English (EN)	Spanish (ES)	Italian (IT)
Train	3200	2646	3200
Validation	800	661	800
Test	1000	831	1000
MTR_{train} (%)	44.6	49.9	45.7
MTR_{test} (%)	46.0	49.9	50.9
Source	Evalita2018	IberEval2018	Evalita2018

Multilingual Sentiment Lexicon[5] Since hate speech often expresses more negative sentiments [25], we utilise a sentiment lexicon, which consists of positive and negative words in 136 languages, and provides 2,955 negative words for English, 2,720 for Spanish and 2,893 for Italian [8].

4.3 Baselines

We compare both CCNL and CCNL-Ex (with lexicons infused) with ten baselines, including SVM with unigrams features, CNN, BiLSTM and CapsNet with monolingual Fasttext embeddings of translated target data, multilingual embeddings MUSE and LASER fed to a 2-layer feedforward neural network, the state-of-the-art cross-lingual models mBERT and XLM-R covered by a 2-layer feedforward classifier on the output layer, and hate-specific cross-lingual model JL-HL with two inputs proposed by Pamungkas and Patti [28]. All baselines are described as follows:

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

SVM. Support Vector Machine (SVM) aims to determine the best decision boundary between vectors that belong to a given category or not [12];

CNN. Consisting of one convolutional layer and one max pooling layer to capture local textual features [23];

BiLSTM. Composed of forward/backward recurrent neural networks to extract long-term dependencies of a text [10];

CapsNet. A single capsule network [31] using a convolutional layer to extract n-gram features;

LASER. Language-Agnostic SEntence Representations (LASER) aims to calculate and use joint multilingual sentence embeddings across 93 languages [2];

MUSE. Multilingual Unsupervised and Supervised Embeddings (MUSE) builds bilingual dictionaries and aligns monolingual word embedding spaces without supervision [24];

mBERT. Multilingual BERT[6] (mBERT) is a variant of BERT [14] that was trained on 104 languages of Wikipedia;

XLM-R. XLM-RoBERTa is a scaled cross-lingual sentence encoder across 100 languages from Common Crawl [11];

---

[1]https://amievalita2018.wordpress.com/data/
[2]https://amiibereval2018.wordpress.com/important-dates/data/
[3]https://translate.google.co.uk/
[4]http://hatespeech.di.unito.it/resources.html

[5]https://sites.google.com/site/datascienceslab/projects/multilingualsentiment
[6]https://github.com/google-research/bert/blob/master/multilingual.md

Table 2: Comparison of CCNL and CCNL-Ex over baselines on the six language pairs. The best result in bold and the second best result underlined.

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

JL-HL. A joint-learning cross-lingual model proposed by Pamungkas and Patti [28], a hybrid approach with LSTM architectures which concatenates multilingual lexical features. The source data and translated target data are fed to two parallels separately.

4.4 Experiment Settings

For training our model, we use FastText embeddings of dimension 300 trained on the Common Crawl and Wikipedia [20]. We use 128 units for forward and backward LSTMs (256 units in total) and 50 units in hidden layer for the feedforward classifier. For capsule networks, we use 10 capsules of dimension 16 and the number of dynamic routing is 5. We use Adam optimiser with 0.0001 learning rate, and set 0.4 for dropout value and 8 for batch size. The model is coded in Keras 2.2.4 and Tensorflow 1.14. We run experiments on the HPC resources of our university, each experiment taking less than one hour. Macro-averaged F1 score is reported as the evaluation metric for all experiments.

5 RESULTS

5.1 Model Performance

Results are shown in Table 2. CCNL and CCNL-Ex differ in that the latter incorporates lexical semantic features. We can observe that CCNL yields better performance than all baseline models for five out of six language pairs, with the exception of ES→EN. CCNL-Ex outperforms all ten baselines for all language pairs. These results substantiate the effectiveness of our model with semantic information, highlighting its generalisation capability across three languages.

Among the ten baselines, the best is JL-HL, whose performance is still always below that of CCNL-Ex. CCNL achieves absolute improvements ranging 7%-9% over JL-HL model for two language pairs involving Italian: EN→IT, and ES→IT. In addition, the CCNL model has manifested pronounced improvements in terms of separately identifying two classes compared to the majority baseline. We can also observe that CCNL generally achieves a large margin with respect to other baselines like SVM, CNN and BiLSTM (especially for ES→EN, EN→IT and IT→ES), while the baseline MUSE achieves good results for IT→EN and IT→ES. It also highlights the

effectiveness of the capsule network as an important component in the cross-lingual model compared with other baselines. Furthermore, we observe that CCNL achieves better performance on all six language pairs when we compare it with CapsNet, LASER, mBERT and XLM-R. Possible reasons are that BiLSTM layers enable the proposed model the capability of extracting contextual information compared to the CNN layer, and CCNL takes the spatial features into consideration by sharing the same weight matrix and learning the positional feature difference in high level via the dynamic routing process.

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

5.2 Comparative Experiments

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

*5.2.1 **Framework Ablation Analysis**. We perform an ablation study for CCNL by dropping one of the two parallel architectures (CCNL-non-parallel), removing the LSTM layer (CCNL-non-LSTM) and removing the Capsule Network layer (CCNL-non-Caps). As shown in Table 3, CCNL outperforms all ablated models, demonstrating the combined benefits of all CCNL components. CCNL noticeably outperforms CCNL-non-parallel on all language pairs, highlighting the importance of the two-parallel framework for extracting local features from both source and target texts. Additionally, we can validate the ability of the BiLSTM network to

Table 3: CCNL comparative experiment results for six language pairs. The best result in bold.

Model	ES→EN	EN→ES	IT→EN	EN→IT	ES→IT	IT→ES
Results for ablation experiments						
CCNL-non-parallel	0.522	0.558	0.570	0.513	0.626	0.624
CCNL-non-LSTM	0.373	0.609	0.565	0.406	0.685	0.623
CCNL-non-Caps	0.597	0.678	0.613	0.439	0.643	0.622
CCNL	**0.624**	**0.719**	**0.628**	**0.584**	**0.737**	**0.668**
Results for feature layers						
CCNL-non-FE	0.373	0.609	0.565	0.406	0.685	0.623
CCNL-CNN	0.521	0.592	0.577	0.439	0.633	0.622
CCNL-GRU	0.458	**0.722**	0.613	0.411	0.715	**0.671**
CCNL	**0.624**	0.719	**0.628**	**0.584**	**0.737**	0.668

Table 4: Examples for error analysis. Translated texts are presented for non-English instances. Ground truth (GT) and prediction (P) labels are noted – hateful (1) and non-hateful (0), along with corresponding error types (ET).

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

extract contextual information effectively compared with CCNL-non-LSTM, which highlights the effectiveness of the capsule network compared with CCNL-non-Caps.

*5.2.2 **Impact of Feature Extraction Layer**.* We aim to validate the ability of the BiLSTM network to extract contextual information effectively. We test different feature extraction layers by keeping other components of the CCNL architecture unchanged. We test four flavours of CCNL: CCNL (with LSTM feature extraction layer), CCNL-non-FE (without feature extraction layer), CCNL-CNN (CNN feature extraction instead) and CCNL-GRU (bidirectional GRU feature extraction instead). Results in Table 3 show that CCNL with the feature extraction layer performs consistently better than those without it, highlighting the importance of extracting local features from the text. Additionally, CCNL also shows improved performance on all tasks when compared with CCNL-CNN, since contextual information plays a significant role in detecting hate speech. CCNL noticeably outperforms CCNL-GRU on four out of six language pairs and achieves similar performance (differences below 0.5%) on the other two language pairs (EN→ES and IT→ES). This is likely due to structural similarities of GRU and LSTM, with the additional complexity of LSTM allowing to capture more informative features in some language pairs.

5.3 Error Analysis

We inspect frequent errors across misclassifications from the test set by the CCNL-Ex model (see Table 4 for examples). We summarise the following four main types of errors:

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

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

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

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

6 CONCLUSION AND FUTURE WORK

We propose a Cross-lingual Capsule Network Learning model integrating Extra hate-related semantic features (CCNL-Ex) for hate speech detection. CCNL, main framework of our model, is composed of two parallel architectures for source and target languages, using BiLSTM to extract contextual features and Capsule Network to capture hierarchically positional relationships. Our model finally leads to state-of-the-art performance for all six language pairs compared with ten competitive baselines. Results show the potential of learning contextual information and spatial relationships of hate speech texts. Given that using machine translation resources such as Google translate is not always a perfect option with very informal language such as those used on social media for hate speech, we will explore approaches to consider culturally grounded context

for parallel corpora in future work. Expansive implementations for more languages and datasets are also desired to analyse the generalisability of our model.

ACKNOWLEDGMENTS

REFERENCES

[1] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2020. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems* (2020), 101584. https://doi.org/10.1016/j.is.2020.101584

[2] Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7 (2019), 597–610. https://doi.org/10.1162/tacl_a_00288

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

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

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

[6] Erik Cambria, Robyn Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI Fall Symposium Series*.

[7] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3098–3112. https://doi.org/10.18653/v1/P19-1299

[8] Yanqing Chen and Steven Skiena. 2014. Building Sentiment Lexicons for All Major Languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 383–389. https://doi.org/10.3115/v1/P14-2063

[9] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 251–262. https://doi.org/10.1145/3308558.3313600

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[13] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[15] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 1606–1615. https://doi.org/10.3115/v1/N15-1184

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

[17] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018.. In *IberEval@ SEPLN*. 214–228.

[18] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 85.

[19] Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6350–6365. https://doi.org/10.18653/v1/2020.coling-main.559

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

[21] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

[22] Geoffrey E. Hinton, Sara Sabour, and Nicholas Frosst. 2018. Matrix capsules with EM routing. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HJWLfGWRb

[23] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[24] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=H196sainb

[25] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712* (2018).

[26] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4675–4684. https://doi.org/10.18653/v1/D19-1474

[27] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management* 57, 6 (2020), 102360. https://doi.org/10.1016/j.ipm.2020.102360

[28] Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and Cross-lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 363–370. https://doi.org/10.18653/v1/P19-2051

[29] Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5838–5844. https://doi.org/10.18653/v1/2020.emnlp-main.470

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

[31] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic Routing Between Capsules. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 3856–3866. https://proceedings.neurips.cc/paper/2017/hash/2cad8fa47bbef282badbb8de5374b894-Abstract.html

[32] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, 1–10. https://doi.org/10.18653/v1/W17-1101

[33] Saurabh Srivastava and Prerna Khurana. 2019. Detecting Aggression and Toxicity using a Multi Dimension Capsule Network. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 157–162. https://doi.org/10.18653/v1/W19-3517

[34] Saurabh Srivastava, Prerna Khurana, and Vartika Tewari. 2018. Identifying Aggression and Toxicity in Comments using Capsule Network. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Association for Computational Linguistics, Santa Fe, New Mexico, USA, 98–105. https://www.aclweb.org/anthology/W18-4412

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[36] Bertie Vidgen and Leon Derczynski. 2021. Directions in Abusive Language Training Data: Garbage In, Garbage Out. *PLOS ONE* 15, 12 (2021), 1–32. https://doi.org/10.1371/journal.pone.0243300

[37] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, 138–142. https://doi.org/10.18653/v1/W16-5618

[38] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

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

[40] Wei Zhao, Jianbo Ye, Min Yang, Zeyang Lei, Suofei Zhang, and Zhou Zhao. 2018. Investigating Capsule Networks with Dynamic Routing for Text Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3110–3119. https://doi.org/10.18653/v1/D18-1350

[41] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1403–1412. https://doi.org/10.18653/v1/P16-1133