Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Tâches Auxiliaires Multilingues pour le Transfert de Modèles de Détection de Discours Haineux

Arij Riabi* Syrielle Montariol* Djamé Seddah
INRIA Paris, F-75012 Paris, France
prenom.nom@inria.fr

RÉSUMÉ

La tâche de détection de contenus haineux est ardue, car elle nécessite des connaissances culturelles et contextuelles approfondies ; les connaissances nécessaires varient, entre autres, selon la langue du locateur ou la cible du contenu. Or, des données annotées pour des domaines et des langues spécifiques sont souvent absentes ou limitées. C'est là que les données dans d'autres langues peuvent être exploitées ; mais du fait de ces variations, le transfert cross-lingue est souvent difficile. Dans cet article, nous mettons en évidence cette limitation pour plusieurs domaines et langues et montrons l'impact positif de l'apprentissage de tâches auxiliaires multilingues - analyse de sentiments, reconnaissance des entités nommées et tâches reposant sur des informations morpho-syntaxiques - sur le transfert cross-lingue zéro-shot des modèles de détection de discours haineux, afin de combler ce fossé culturel.

ABSTRACT

Multilingual Auxiliary Tasks for Zero-Shot Cross-Lingual Transfer of Hate Speech Detection Models.

Detecting hateful content is a challenging task, as it requires extensive cultural and contextual knowledge from the model ; the necessary knowledge varies depending on the speaker's language or the target of the content. However, annotated data for specific domains and languages are often inexistant or limited. In that case, annotated data in other languages can be exploited ; but the cross-lingual transfer is often difficult due to these cultural and contextual variations. In this paper, we highlight this limitation for several domains and languages and show the positive impact of learning multilingual auxiliary tasks - sentiment analysis, recognition, and tasks based on morpho-syntactic information - on the cross-lingual zero-shot transfer of hate speech detection models in order to bridge this cultural gap.

MOTS-CLÉS : Détection des Discours Haineux, Transfert Cross-lingue, Apprentissage multitâche.

KEYWORDS: Hate Speech Detection, Cross-lingual Transfert, Multi-task learning.

1 Introduction

Étant donné l'impact et les enjeux considérables, ne serait-ce qu'en terme de réorientation du discours politique, liés à la diffusions de contenu haineux, en particulier sur les médias sociaux, il est naturel, voire heureux, que la communauté consacre des efforts considérables à la détection de ce type de contenus. Cela passe par la proposition de différents type d'approches, qu'elles soient basées sur des règles manuscrites ou sur de l'apprentissage automatique, et le déploiement de *benchmarks* et jeux

∗. Ces auteurs ont contribué de manière égale.

Volume 1 : conférence principale, pages 413–423.

de données d'évaluation permettant d'évaluer leurs performances et plus crucialement, leurs limites (Valette, 2004; Sood *et al.*, 2012; Waseem & Hovy, 2016; Papegnies *et al.*, 2017; Davidson *et al.*, 2017; Fortuna & Nunes, 2018; Kennedy *et al.*, 2020; Zampieri *et al.*, 2021).

Il est important de noter que les systèmes orientés données sont conçus pour être efficaces à un moment donné et pour un type spécifique de contenu en ligne sur lequel ils ont été entraînés. Or les discours haineux varient considérablement selon la période ou la culture du locuteur (Florio *et al.*, 2020). Par exemple, le renforcement de la censure des plateformes en ligne force les utilisateurs de médias sociaux à s'adapter à leurs systèmes de filtrage, en faisant évoluer leur lexique et en introduisant des variations orthographiques, leur évitant ainsi d'être détectés (Berger & Perez, 2016; Vidgen *et al.*, 2019). Des événements spécifiques peuvent en outre déclencher l'apparition de nouveaux domaines de discours haineux, conduisant à de nouveaux lexiques, hashtags et marqueurs pictographiques qui de facto rendent les corpus d'entraînement inadaptés voire quasiment obsolètes (Markov *et al.*, 2021).

De manière synchronique, les variations culturelles modifient la perception et la forme des discours de haine d'une langue à l'autre. Ainsi, certaines insultes peuvent être considérées comme non offensantes dans une langue – dénotant tout de même un registre informel – mais seront considérées comme offensantes, voire haineuses, dans une autre langue (Nozza, 2021). Dans cet article, nous étudions les variations cross-lingues des discours de haine en tant que *fossé culturel*. [1]

Dans des scénarios de détection de discours haineux (DDH) où les ressources linguistiques sont limitées, plusieurs travaux proposent des méthodes pour transférer les modèles de détection d'une langue à l'autre (Basile & Rubagotti, 2018; van der Goot *et al.*, 2018; Pamungkas & Patti, 2019; Ranasinghe & Zampieri, 2020; Chiril *et al.*, 2019; Mozafari *et al.*, 2022). Cependant, les variations culturelles et linguistiques des discours de haine peuvent réduire la transférabilité de ces modèles. L'information sur ces variations spécifique aux données cibles monolingues du discours haineux doit être trouvée ailleurs dans notre cadre de "zéro-shot".

Nous émettons l'hypothèse que ces informations peuvent être capturées en *fine-tunant* le modèle de langue sur des tâches riches en ressources à la fois dans la langue source et dans la langue cible du transfert (van der Goot *et al.*, 2021a). Ainsi, en *fine-tunant* un modèle multilingue (XLM-R, (Conneau *et al.*, 2020)) avec une architecture multitâche, nous étudions l'impact de tâches auxiliaires opérant à différents niveaux de granularité linguistique [2] sur l'efficacité du transfert. À partir d'un jeu de tweets annotés en contenu haineux dans deux domaines (discours de haine contre les femmes et anti-immigrants) et trois langues (anglais, italien et espagnol), nous construisons un cadre expérimental strictement comparable et mettons en évidence le problème du *fossé culturel* pour le transfert cross-lingue zéro-shot des modèles de détection des discours haineux. Dans le cadre de la détection des discours haineux, notre travail démontre l'impact positif de l'entraînement conjoint de tâches auxiliaires sur le transfert cross-lingue, notamment les tâches d'analyse de sentiments et de NER.

2 Travaux connexes

Amélioration des modèles de langue à l'aide d'autres tâches. Afin d'améliorer l'efficacité d'un modèle de langue pré-entraîné pour une tâche donnée, ce modèle peut être préalablement *fine-tuné*

1. Cette notion est également introduite par Cabrio *et al.* (2014), qui étudie l'écart entre les langues dans les entrées de DBpedia.

2. Etiquetage morpho-syntaxique, reconnaissance d'entités nommées, analyse en dépendances syntaxique et analyse des sentiments.

sur une tâche intermédiaire (Pruksachatkun *et al.*, 2020). Cette idée a été formalisée par Phang *et al.* (2018) avec un pré-entraînement séquentiel monolingue. Dans un cadre cross-lingue, Phang *et al.* (2020) *fine-tunent* un modèle sur neuf tâches intermédiaires de compréhension de l'anglais, et l'appliquent à un ensemble de tâches cible dans une langue différente tandis que nous utilisons des données auxiliaires dans toutes les langues. Ils montrent que l'efficacité des systèmes diffère beaucoup en fonction des tâches.

Dans le contexte de la détection de contenus haineux, l'entraînement sur des tâches auxiliaires est fait presque exclusivement sur de l'analyse de sentiments (Bauwelinck, Nina and Lefever, Els, 2019; Aroyehun & Gelbukh, 2021) et uniquement dans des scénarios monolingues. Mais des informations supplémentaires sont parfois ajoutées au classificateur de manière plus directe. Gambino & Pirrone (2020), un des meilleurs systèmes sur la tâche HaSpeeDe de l'EVALITA 2020, utilise du texte étiqueté morpho-syntaxiquement en entrée de ses modèles. L'information syntaxique est également ajoutée à des plongements non contextuels par Narang & Brew (2020) en utilisant des classificateurs basés sur la structure syntaxique du texte avec un effet positif sur la détection de langage abusif.

Apprentissage par transfert cross-lingue zéro-shot pour la détection de contenus haineux. En raison du manque de données annotées sur de nombreuses langues et domaines , le transfert cross-lingue zéro-shot a été beaucoup abordé dans la littérature. Parmi les travaux les plus récents, Pelicon *et al.* (2021) étudient l'impact d'un entraînement d'un modèle de classification sur des données de discours haineux dans des langues différentes de la langue cible. Ils montrent que les modèles de langue pré-entraînés sur un faible nombre de langues bénéficient davantage de cet entraînement intermédiaire et surpassent souvent les modèles de langue multilingues massifs. Nozza (2021), sur lequel s'appuie notre travail, démontre la limite du transfert cross-lingue pour les discours de haine spécifiques à un domaine – en particulier, les discours de haine envers les femmes – et l'explique en montrant des exemples de variation culturelle entre les langues. Par exemple, certains termes de discours haineux notables dans une langue peuvent être utilisés comme intensificateur dans une autre langue. [3]

3 Un transfert zéro-shot cross-lingue limité

Nous partons des mêmes ensembles de données de discours haineux que Nozza (2021), qui s'en est servi pour souligner les limites du transfert cross-lingue zéro-shot. Les corpus sont en trois langues : anglais (en), espagnol (es) et italien (it). Chaque langue est divisée en corpus provenant de deux domaines : discours de haine envers les immigrants et discours de haine envers les femmes. Il s'agit d'une tâche de classification binaire avec deux étiquettes : *haineux* et *non haineux*.

Corpus comparables. Contrairement aux expériences de Nozza (2021), nous construisons des corpus comparables dans chaque langue et domaine afin de garantir la comparabilité des paramètres de transfert, en réduisant tous les ensembles de données à une taille totale de 2 591 tweets, soit la taille du plus petit d'entre eux. Chaque ensemble d'entraînement compte 1 618 tweets, chaque ensemble de développement 173, et chaque ensemble de test 800. [4] La répartition des classes se trouve dans la

3. Nozza (2021) donne l'exemple du mot espagnol *puta* souvent utilisé comme intensificateur sans aucune connotation misogyne, alors qu'il se traduit par une version argotique de *prostituée* en anglais.

4. Lorsque nous sous-échantillonnons les corpus, nous nous assurons qu'ils restent comparables : nous utilisons le test de Kolmogorov–Smirnov pour comparer la distribution de la longueur des phrases (nombre de mots) et le pourcentage de discours haineux entre les ensembles de données échantillonnés et les ensembles de données originaux.

Table 1. Les classes sont relativement balancées, l'étiquette 'haineux' étant en général légèrement sous-représentée.

Langue	immigrants	femmes
Anglais	39.36%	43.75%
Italien	40.66%	40.97%
Espagnol	33.00%	46.04%

TABLE 1 – Pourcentage des exemples haineux par langue dans l'ensemble d'entraînement.

Pré-traitement et entraînement. Nous préparons les corpus en remplaçant toutes les mentions et les urls par des tokens spécifiques, et en segmentant les hashtags en mots. Étant donné la nature compositionnelle des hashtags (un ensemble de mots concaténés), la segmentation est fréquemment effectuée comme étape de traitement dans la littérature lors de l'analyse de tweets (Röttger *et al.* (2021)). Cette étape peut améliorer des tâches telles que le regroupement de tweets (Gromann & Declerck, 2017). Ce point constitue une différence importante avec les travaux de Nozza (2021).

Nous *fine-tunons* un modèle de langue multilingue pré-entraîné sur la tâche de DDH pour chaque corpus, dans les deux domaines et trois langues. Ensuite, nous appliquons le modèle de classification binaire entraîné sur l'ensemble de test de chaque langue cible en étudiant deux scénarios : (i) monolingue, c.-à-d. entraînement et test sur la même langue et le même domaine pour les discours haineux ; (ii) zéro-shot cross-lingue, c.-à-d. entraînement sur une langue et test sur une autre.

Les ensembles de test échantillonnés à partir des corpus originaux sont de taille relativement limitée (800 observations). Pour augmenter la robustesse des résultats, nous utilisons cinq graines aléatoires différentes lors de l'entraînement et nous rapportons la macro-F1 moyenne sur les cinq entraînements. Nous conservons la meilleure des 20 époques pour chaque entraînement en fonction du score macro-F1 sur l'ensemble de développement.

Modèle de référence pour la détection de contenu haineux. Nous rapportons les résultats de référence avec le modèle XLM-R (Conneau *et al.*, 2020) dans la Table 2. On remarque que les modèles entraînés et testés sur le corpus anglais dans le domaine des immigrants ont des scores particulièrement bas. Ce phénomène a également été observé dans la littérature (Nozza, 2021; Stappen *et al.*, 2020), et s'explique par la présence dans l'ensemble d'entraînement de mots et hashtags spécifiques ayant été utilisés pour obtenir les tweets et qui conduisent le modèle à sur-apprendre.

Pour résumer les résultats, nous les agrégeons selon les deux scénarios présentés ci-dessus : monolingue (*mono*), et cross-lingue zéro-shot (*cross*). Pour chaque domaine (immigrants et femmes), nous moyennons les scores par scénario : les colonnes "mono" montrent la moyenne de tous les scores en italique dans le tableau de gauche tandis que la colonne "cross" est la moyenne des résultats pour tous les autres scores. Nous observons un phénomène problématique lors du transfert : pour le domaine *femmes*, les modèles testés sur les ensembles de test en espagnol et italien en mode zéro-shot ont des scores beaucoup plus faibles que pour le domaine *immigrants* (les chiffres en orange dans le tableau). Cela est dû, comme le montre Nozza (2021), aux interjections spécifiques à la langue qui conduisent le modèle à classer à tort certains textes comme haineux envers les femmes.

Langue source	immigrants			femmes		
	en	es	it	en	es	it
en	*52.8*	44.2	64.6	*49.4*	46.6	44.7
es	68.0	*75.1*	64.9	47.1	*60.0*	43.8
it	64.4	44.2	*74.5*	53.3	52.6	*83.7*

immigrants		femmes	
mono	cross	mono	cross
67.5 (10.38)	58.4 (10.09)	64.4 (14.34)	48.0 (3.67)

TABLE 2 – Résultats de référence (macro-F1 (%) moyennés sur 5 entraînements) avec XLM-R, détaillée (avec les cas monolingues en *italic*) et agrégées (avec l'écart-type des scores en gris). Les chiffres en orange désignent les cas où le transfert cross-lingue zéro-shot échoue.

4 Entraînement sur des tâches auxiliaires

Nous étudions l'impact de l'entraînement sur des tâches auxiliaires pour le problème de transfert zéro-shot cross-lingue des modèles de détection de contenus haineux. En utilisant des données pour les tâches auxiliaires à la fois dans la langue source et la langue cible, nous nous attendons à ce que l'entraînement sur ces tâches fonctionne comme un pont entre les deux langues, aidant le transfert cross-lingue en fournissant plus d'informations sur la langue cible et la différence entre les deux langues.

Nous définissons plusieurs tâches dont nous voulons évaluer l'effet sur le transfert cross-lingue ; une tâche au niveau de la séquence, l'analyse des sentiments, et plusieurs tâches au niveau du mot : la reconnaissance d'entités nommées (NER) et un ensemble de tâches syntaxiques que nous regroupons – par abus de langage – sous le terme de "Universal Dependencies" (UD). Les tâches de NER et d'analyse de sentiments permettent au modèle d'apprendre des informations sémantiques tandis que les tâches UD transmettent des compétences syntaxiques au modèle.

Tâches syntaxiques. Nous étudions l'effet de l'ajout d'informations syntaxiques en utilisant toutes les tâches de dépendances universelles (UD, (Nivre *et al.*, 2020)) (analyse syntaxique de la dépendance, étiquetage morpho-syntaxique (POS), lemmatisation et étiquetage morphologique). Nous utilisons les jeux de données EWT (Silveira *et al.*, 2014), GSD et ISDT (Bosco *et al.*, 2014), pour l'anglais, l'espagnol et l'italien respectivement. Les jeux de données étant de tailles différentes, nous les échantillonnons pour obtenir la même taille d'entraînement dans toutes les langues. Nous utilisons un ensemble d'entraînement de 12 543 phrases, soit la taille de l'ensemble de données le plus petit.

Analyse des sentiments. Nous utilisons des ensembles de données d'analyse de sentiments extraits de Twitter pour chacune de nos trois langues cibles. [5] Ils ont été rassemblés et unifiés par Barbieri *et al.* (2021) avec des tailles identiques (entraînement 1 839, développement 324, test 870) et une distribution équilibrée entre les trois étiquettes de sentiment (positif, négatif et neutre).

Reconnaissance d'entités nommées (NER). Nous utilisons le jeu de données NER WikiANN de (Pan *et al.*, 2017; Rahimi *et al.*, 2019), qui couvre nos trois langues. Les corpus pour chaque langue ont aussi des taille unifiées (entraînement 20 000 exemples, développement 10 000, test 10 000).

5. https://github.com/cardiffnlp/xlm-t

5 Résultats et discussion

Nous effectuons l'apprentissage multitâche en utilisant le système MACHAMP (van der Goot *et al.*, 2021b). Nous *fine-tunons* XLM-R sur toutes les combinaisons des tâches auxiliaires conjointement avec la tâche de DDH, notre tâche cible. Ainsi, pour une tâche auxiliaire donnée, par exemple l'analyse de sentiment, nous fusionnons les corpus de sentiment des trois langues. Puis nous y ajoutons le corpus d'entraînement de la tâche de DDH dans une seule langue (e.g. l'anglais) et entraînons le modèle sur ces données. Pour finir, le modèle entraîné est appliqué en inférence sur les corpus de test des trois langues.

Nous montrons les résultats agrégés en calculant les deltas entre le système de référence (aucune tâche auxiliaire) et le système enrichi avec un entraînement sur tâches auxiliaires (Table 3). Pour chaque domaine (immigrants et femmes), nous faisons la moyenne des deltas pour les cas monolingue et cross-lingue zéro-shot, comme dans la Table 2.

Tâche Auxiliaire	immigrants		femmes	
	mono	cross	mono	cross
XLM-R				
None	67.5 (10.38)	58.4 (10.09)	64.4 (14.34)	48.0 (3.67)
UD	2.9 (4.87)	-2.5 (3.79)	1.8 (2.80)	-3.2 (8.10)
NER	0.9 (7.79)	3.1 (4.05)	2.8 (2.04)	1.7 (5.55)
sentiment	3.0 (5.27)	3.7 (3.85)	1.7 (1.87)	-2.0 (6.18)
sentiment+UD	4.0 (4.23)	5.1 (4.84)	1.1 (1.58)	-2.0 (6.96)
sentiment+NER	0.2 (9.44)	3.7 (1.68)	3.3 (2.92)	-2.0 (2.98)
sentiment+UD+NER	-0.1 (10.62)	5.2 (4.05)	2.8 (1.56)	-2.6 (6.36)
UD+NER	-1.0 (10.25)	-1.6 (11.48)	2.0 (0.94)	-6.5 (8.05)
XLM-T				
None	67.8 (15.80)	60.3 (4.30)	68.1 (11.01)	53.7 (7.67)
UD	2.8 (2.13)	-0.2 (3.19)	-1.1 (2.07)	-8.2 (6.12)
NER	1.2 (4.05)	3.4 (5.17)	1.7 (2.44)	-0.6 (4.30)
sentiment	1.1 (2.70)	2.4 (3.66)	0.1 (2.85)	-0.5 (5.01)
sentiment+NER	0.7 (4.66)	3.8 (4.38)	2.4 (2.99)	-1.5 (3.60)
sentiment+UD	2.1 (1.99)	1.6 (3.75)	0.6 (1.35)	-2.2 (2.92)
sentiment+UD+NER	0.2 (5.85)	2.9 (3.90)	1.4 (2.76)	-2.9 (2.39)
UD+NER	-0.7 (6.59)	1.7 (3.92)	-0.3 (2.79)	-9.1 (8.71)

TABLE 3 – Scores Macro-F1 (%) agrégés des modèles de référence XLM-R et XLM-T, et deltas entre le score de chaque modèle (XLM-R ou XLM-T) entraîné sur les tâches auxiliaires et le score de référence du modèle. Les scores en vert indiquent les cas où les tâches auxiliaires ont un impact positif sur la détection de discours haineux. *Sent* correspond à la tâche d'analyse de sentiment. L'écart-type des scores est indiqué en gris entre parenthèses à côté des deltas.)

L'entraînement sur des tâches auxiliaires améliore globalement la performance pour notre tâche cible monolingue dans les deux domaines (Table 3, colonnes *mono*). Dans le scénario de transfert cross-lingue zéro-shot (*cross*), pour le domaine *immigrants*, les tâches d'analyse de sentiment et NER améliorent les résultats tandis que les tâches UD les dégradent. Ceci est conforme aux résultats de la littérature qui s'accordent sur l'effet positif sur l'analyse des sentiments (del Arco *et al.*, 2021; Aroyehun & Gelbukh, 2021), mais font face à des conclusions variables lorsqu'il s'agit des tâches UD (Narang & Brew, 2020; Klemen *et al.*, 2020; Pruksachatkun *et al.*, 2020). Nous supposons que

lorsque nous travaillons sur des tweets, leur style contraint - des phrases courtes, généralement de faible complexité syntaxique - rend les connaissances syntaxiques supplémentaires inutiles pour une tâche en aval telle que la DDH, qui bénéficie davantage des informations sémantiques.

Les cas avec fossé culturel (contenus ciblant les femmes) restent problématiques. En utilisant XLM-R, la tâche auxiliaire NER conduit à la meilleure amélioration pour ces situations. En effet, la transférabilité cross-lingue de cette tâche est facilitée par le fait que de nombreuses entités nommées sont les mêmes d'une langue à l'autre (par exemple, les noms de personnes et d'organisations).

L'autre méthode efficace pour améliorer le transfert cross-lingue zéro-shot consiste à utiliser un modèle de langue entraîné sur les données Twitter. Nous utilisons XLM-T (Barbieri *et al.*, 2021), un modèle XLM-R entrainé sur 200 millions de tweets collectés entre 2018 et 2020, dans plus de 30 langues, dont nos trois langues cibles. Comme prévu, le modèle obtient globalement de bien meilleurs scores que le modèle XLM-R (Table 3). Ceci est en accord avec les résultats de Bose *et al.* (2021); Muller *et al.* (2021) et van der Goot *et al.* (2021a), qui montrent la supériorité du *Masked Language Modeling* sur les autres tâches dans un cadre de transfert cross-corpus et cross-lingue. Cependant, l'impact des tâches auxiliaires sur les performances de DDH est comparable à celui observé avec XLM-R : positif partout sauf pour le transfert cross-lingue zéro-shot pour les contenus haineux ciblant les femmes.

Dans ce dernier cas, le problème du fossé culturel doit être distingué de la variation entre les ensembles de données. Le contenu, le sujet et le vocabulaire des tweets peuvent différer considérablement d'un corpus de discours haineux à l'autre (certains événements et comptes ont été spécifiquement ciblés lors de l'extraction de tweets, par exemple les victimes du Gamergate pour les ensembles de données italiens sur les discours haineux à l'égard des femmes (Fersini *et al.*, 2018)). Des disparités sont aussi présentes vis-à-vis des données d'entraînement de XLM-T et des corpus d'analyse de sentiments de Twitter. De fait, le modèle XLM-T est uniquement adapté au style et à la forme des contenus Twitter, et est lié à la période de collecte et aux domaines que ses données d'entraînement recoupent.

6 Conclusion

Nous avons montré l'impact positif d'entraîner conjointement la détection des discours haineux avec des tâches de NER et d'analyse de sentiments. Nous avons mis en évidence des cas problématiques où le transfert cross-lingue zéro-shot, pour les discours de haine ciblant les femmes, échoue. Dans cette situation, l'hypothèse émise dans la littérature est qu'un fossé culturel entre les langues limite la transférabilité du modèle (Nozza, 2021) ; cette variation serait liée, par exemple, à l'usage de certains qualificatifs associés à des femmes qui seraient considérés comme offensants dans une langue et pas dans une autre.

Dans ces cas, nous identifions les deux solutions les plus efficaces : l'entraînement auxiliaire sur la tâche NER et l'utilisation d'un modèle de langue entraîné sur des données plus adaptées (ici, des tweets), qui permettent au modèle de discerner des schémas communs entre les langues source et cible et de transférer des connaissances de l'une à l'autre.

Remerciements

Ce travail a reçu le financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne sous la convention de subvention numéro 101021607. Le dernier auteur a reçu le soutien de l'Agence française de la recherche via le projet ANR ParSiTi (ANR16-CE33-0021).

Références

AROYEHUN S. T. & GELBUKH A. (2021). Evaluation of intermediate pre-training for the detection of offensive language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings. CEUR-WS. org*.

BARBIERI F., ESPINOSA-ANKE L. & CAMACHO-COLLADOS J. (2021). A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv :2104.12250*.

BASILE A. & RUBAGOTTI C. (2018). Crotonemilano for ami at evalita2018. a performant, cross-lingual misogyny detection system. *EVALITA Evaluation of NLP and Speech Tools for Italian*, **12**, 206.

BAUWELINCK, NINA AND LEFEVER, ELS (2019). Measuring the impact of sentiment for hate speech detection on Twitter. In FOLDS, DENNIS AND LEFEVER, ELS AND GERA, RALUCCA AND HOSTE, VERONIQUE, Éd., *Proceedings of HUSO 2019, The fifth international conference on human and social analytics*, p. 17–22 : IARIA, International Academy, Research, and Industry Association.

BERGER J. & PEREZ H. (2016). *The Islamic State's diminishing returns on Twitter : How suspensions are limiting the social networks of English-speaking ISIS supporters*. Rapport interne, George Washington University.

BOSCO C., DELL'ORLETTA F., MONTEMAGNI S., SANGUINETTI M. & SIMI M. (2014). The evalita 2014 dependency parsing task. *The Evalita 2014 Dependency Parsing task*, p. 1–8.

BOSE T., ILLINA I. & FOHR D. (2021). Unsupervised domain adaptation in cross-corpora abusive language detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, p. 113–122, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.socialnlp-1.10.

CABRIO E., COJAN J. & GANDON F. (2014). Mind the cultural gap : Bridging language-specific dbpedia chapters for question answering. In *Towards the Multilingual Semantic Web*, p. 137–154. Springer.

CHIRIL P., BENAMARA F., MORICEAU V., COULOMB-GULLY M. & KUMAR A. (2019). Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019)*, p. 351–360 : ATALA.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DAVIDSON T., WARMSLEY D., MACY M. & WEBER I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

DEL ARCO F. M. P., HALAT S., PADÓ S. & KLINGER R. (2021). Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. In *Forum for Information Retrieval Evaluation, Virtual Event*.

FERSINI E., NOZZA D. & ROSSO P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, **12**, 59.

FLORIO K., BASILE V., POLIGNANO M., BASILE P. & PATTI V. (2020). Time of your hate : The challenge of time in hate speech detection on social media. *Applied Sciences*, **10**, 4180. DOI : 10.3390/app10124180.

FORTUNA P. & NUNES S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, **51**(4), 1–30.

GAMBINO G. & PIRRONE R. (2020). Chilab@ haspeede 2 : Enhancing hate speech detection with part-of-speech tagging. -.

GROMANN D. & DECLERCK T. (2017). Hashtag processing for enhanced clustering of tweets. In *RANLP*, p. 277–283.

KENNEDY B., JIN X., MOSTAFAZADEH DAVANI A., DEHGHANI M. & REN X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5435–5442, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.483.

KLEMEN M., KRSNIK L. & ROBNIK-ŠIKONJA M. (2020). Enhancing deep neural networks with morphological information. *arXiv preprint arXiv :2011.12432*.

MARKOV I., LJUBEŠIĆ N., FIŠER D. & DAELEMANS W. (2021). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 149–159, Online : Association for Computational Linguistics.

MOZAFARI M., FARAHBAKHSH R. & CRESPI N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*.

MULLER B., ANASTASOPOULOS A., SAGOT B. & SEDDAH D. (2021). When being unseen from mBERT is just the beginning : Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 448–462, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.38.

NARANG K. & BREW C. (2020). Abusive language detection using syntactic dependency graphs. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, p. 44–53, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.alw-1.6.

NIVRE J., DE MARNEFFE M.-C., GINTER F., HAJIČ J., MANNING C. D., PYYSALO S., SCHUSTER S., TYERS F. & ZEMAN D. (2020). Universal Dependencies v2 : An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France : European Language Resources Association.

NOZZA D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 907–914, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-short.114.

PAMUNGKAS E. W. & PATTI V. (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, p. 363–370, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-2051.

PAN X., ZHANG B., MAY J., NOTHMAN J., KNIGHT K. & JI H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1946–1958, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1178.

PAPEGNIES E., LABATUT V., DUFOUR R. & LINARÈS G. (2017). Detection of abusive messages in an on-line community. In *CORIA*, p. 153–168.

PELICON A., SHEKHAR R., ŠKRLJ B., PURVER M. & POLLAK S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, **7**, e559.

PHANG J., CALIXTO I., HTUT P. M., PRUKSACHATKUN Y., LIU H., VANIA C., KANN K. & BOWMAN S. R. (2020). English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, p. 557–575, Suzhou, China : Association for Computational Linguistics.

PHANG J., FÉVRY T. & BOWMAN S. R. (2018). Sentence encoders on stilts : Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv :1811.01088*.

PRUKSACHATKUN Y., PHANG J., LIU H., HTUT P. M., ZHANG X., PANG R. Y., VANIA C., KANN K. & BOWMAN S. R. (2020). Intermediate-task transfer learning with pretrained language models : When and why does it work ? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5231–5247, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.467.

RAHIMI A., LI Y. & COHN T. (2019). Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 151–164, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1015.

RANASINGHE T. & ZAMPIERI M. (2020). Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5838–5844, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.470.

RÖTTGER P., VIDGEN B., NGUYEN D., WASEEM Z., MARGETTS H. & PIERREHUMBERT J. (2021). HateCheck : Functional tests for hate speech detection models. 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)*, p. 41–58, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.4.

SILVEIRA N., DOZAT T., DE MARNEFFE M.-C., BOWMAN S., CONNOR M., BAUER J. & MANNING C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2897–2904, Reykjavik, Iceland : European Language Resources Association (ELRA).

SOOD S., ANTIN J. & CHURCHILL E. (2012). Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, p. 1481–1490, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/2207676.2208610.

STAPPEN L., BRUNN F. & SCHULLER B. (2020). Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel. *arXiv preprint arXiv :2004.13850*.

VALETTE M. (2004). Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur internet. In *Colloque International sur le Document Electronique*, p. 215–230 : Centre de recherche en Ingénierie Multilingue, INaLCO.

VAN DER GOOT R., LJUBEŠIĆ N., MATROOS I., NISSIM M. & PLANK B. (2018). Bleaching text : Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 383–389, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-2061.

VAN DER GOOT R., SHARAF I., IMANKULOVA A., ÜSTÜN A., STEPANOVIĆ M., RAMPONI A., KHAIRUNNISA S. O., KOMACHI M. & PLANK B. (2021a). From masked language modeling to translation : Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2479–2497, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.197.

VAN DER GOOT R., ÜSTÜN A., RAMPONI A., SHARAF I. & PLANK B. (2021b). Massive choice, ample tasks (MaChAmp) : A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 176–197, Online : Association for Computational Linguistics.

VIDGEN B., HARRIS A., NGUYEN D., TROMBLE R., HALE S. & MARGETTS H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, p. 80–93, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3509.

WASEEM Z. & HOVY D. (2016). Hateful symbols or hateful people ? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, p. 88–93, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-2013.

ZAMPIERI N., ILLINA I., FOHR D., D'SA A. G., ILLINA I., FOHR D., KLAKOW D., RUITER D., ILLINA I., FOHR D. *et al.* (2021). A comparative study of different state-of-the-art nlp models for efficient automatic hate speech detection. In *COMMENTS, HATE SPEECH, DISINFORMATION AND PUBLIC COMMUNICATION REGULATION*.