Computational Linguistics in the Netherlands Journal 12 (2022) 141-164 Submitted 12/2022; Published 12/2022

All That Glitters is Not Gold: Transfer-learning for Offensive Language Detection in Dutch

Dion Theodoridis DIONTHEODORIDIS@GMAIL.COM
Tommaso Caselli T.CASELLI@RUG.NL

Center for Language of Cognition, University of Groningen, The Netherlands

Abstract

Creating datasets for language phenomena to fill gaps in the language resource panorama of specific natural languages is not a trivial task. In this work, we explore the application of transfer-learning as strategy to boost both the creation of language-specific datasets and systems. We use offensive language in Dutch tweets directed at Dutch politicians as a case study. In particular, we trained a multilingual model using the Political Speech Project (Bröckling et al. 2018) dataset to automatically annotate tweets in Dutch. The automatically annotated tweets have been used to further train a monolingual language model in Dutch (BERTje) adopting different strategies and combination of manually curated data. Our results show that: (i) transfer learning is an effective strategy to boost the creation of new datasets for specific language phenomena by reducing the annotation efforts; (ii) using a monolingual language model fine-tuned with automatically annotated data (i.e., silver data) is a competitive baseline against the zero-shot transfer of a multilingual model; and finally, (iii) less surprisingly, the addition of automatically annotated data to manually curated ones is a source of errors for the systems, degrading their performances.

Warning: *this paper contains examples of offensive language that may be disturbing, harmuful, and distressing to some readers. Following best practices in this field, slurs and swear words have been obfuscated.*

1. Introduction

The creation of language-specific language resources is costly, in terms of time, money, and human effort. Alternative solutions have seen the application of transfer learning techniques to tackle this problem, especially in cross-lingual settings. We can describe transfer learning as the application of knowledge that a system (or a model) has obtained in one task, using some form of supervised training, to a different one. This may correspond to a different domain, language variety (Ramponi and Plank 2020, Karouzos et al. 2021, Bose et al. 2021), or language (Artetxe and Schwenk 2019, Nooralahzadeh et al. 2020, Huang et al. 2021). From a certain perspective, the fine-tuning of pre-trained language models (PTLMs) on a specific task (e.g., document classification) can also be seen as a form of transfer learning.

In this work, transfer learning, and in particular zero-shot cross-lingual transfer learning, is used as a method to investigate its contribution to the development of language specific datasets and systems. As a case study, we focus on offensive language against politicians in Twitter messages in Dutch. While Dutch has quite a strong Natural Language Processing (NLP) panorama, it is less-resourced when it comes to datasets and tools for socially unacceptable language phenomena.

Socially unacceptable language is a broad term that covers multiple phenomena that have natural language as the primary means of expression, ranging from toxic language and microaggressions, to hate speech, abusive language and offensive language, among others. Socially unacceptable language represents an issue for the full development of inclusive and peaceful societies. The growth and spread of Social Media platforms (e.g., Facebook, Twitter, Reddit, Instagram, 4Chan, among others) has contributed to an increase of this phenomenon as well as to a polarization of societies (Cinelli et al. 2021, Haidt 2022). The mainstream approach in automatic content moderation is still based on so-called reactive interventions, i.e., blocking or deleting the "bad" messages (Seering et al. 2019). Whether this is the most effective strategy to counteract socially unacceptable languages is still a matter of debate (Chandrasekharan et al. 2017), also considering the risks that this approach has in perpetrating bias and discrimination (Sap et al. 2019). Regardless of the preferred and most effective method, the first step is the detection and identification of specific types of socially unacceptable language. Such step necessarily requires language-specific resources to train tools to distinguish the "good" messages from the harmful ones.

In this work, we focus on a specific sub-type of socially unacceptable language, namely offensive messages. We present a series of experiments that aims to assess the contribution of automatically annotated data obtained by applying zero-shot cross-lingual transfer learning using mBERT (Pires et al. 2019) to Twitter messages in Dutch targeting politicians. To better assess the impact and the contributions of zero-shot cross-lingual transfer, we manually annotated a set of 1,500 tweets. This test set has been developed by applying the same guidelines for offensive language used to develop the Dutch Abusive Language Corpus (DALC, henceforth) (Caselli et al. 2021a, Ruitenbeek et al. 2022), thus extending the availability of test data for offensive language in Dutch in the spirit of dynamic benchmarking (Kiela et al. 2021, Thrush et al. 2022).

Our main contributions are:

- the availability of a new benchmark test in Dutch for offensive language on Twitter messages containing the mentions of politicians;
- a set of experiments investigating the contribution of automatically annotated data as strategies to boost the development of competitive NLP systems in absence of manually annotated training data;
- an analysis of the benefits of zero-shot transfer learning to boost the creation of manually curated datasets.

The paper is further structured as follows: an overview of related work on offensive language detection and the use of cross-lingual transfer learning is presented in Section 2. Section 3 describes the data we have collected and used, distinguishing between manually annotated, unlabelled, and automatically annotated. Our experiment settings, the new test benchmark and the results are presented in Section 4, followed by a discussion and critical reflections in Section 5. Finally, in Section 6 conclusions and future work are presented.

2. Related Work

Socially unacceptable language comprehends many different phenomena with varying degree of harmfulness. It is out of the scope of this contribution to present an exhaustive overview of all phenomena, datasets, and machine learning-based approaches related to their treatment (we refer interested readers to Poletto et al. (2021) and Dhanya and Balakrishnan (2021) for recent surveys). On the other hand, we will structure this review along three core topics: offensive language detection, cross-lingual transfer learning, and the impact of using automatically annotated data in low-data settings.

2.1 Offensive Language Detection: Definition, Data, Methods

Offensive language is primarily a subjective phenomenon since it is strictly connected to the perception of a message by the receiver (Vidgen et al. 2019). Defining what is offensive language is a more challenging task than one can imagine. The Merriam-Webster Online Dictionary states that "offensive" corresponds to "*causing displeasure or resentment*",[1] highlighting the negative feelings

[1] `https://www.merriam-webster.com/dictionary/offensive`

that an offence may trigger. Poletto et al. (2017), on the other hand, adopts a definition of offensive language that somehow recalls the (lack of) definition of obscenity by Justice Potter Stewart in *Jacobellis v. Ohio (1964)*:[2] something is offensive as soon as there is even the smallest possibility of being perceived as offensive by someone. Whereas this definition concentrates on the negative feeling that offensiveness may have on others, Zampieri et al. (2019a) broadens the definition by explicitly mentioning non-acceptable language. In particular, in their work offensive language is defined as:

> Posts containing any form of non-acceptable language (profanity) or a targeted offence, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.

While subjectivity has not completely disappeared, this definition is as an attempt to reduce it by introducing more verifiable parameters such as the presence of profanities, swear words, and threats. At the same time, there are risks of introducing bias in datasets based on this definition by reducing offensive messages to explicit realizations only. Besides this limitation, this definition of offensive languages has been adopted in two subsequent SemEval tasks, namely SemEval 2019 Task 6: OffensEval (Zampieri et al. 2019c) (English data only) and SemEval 2020 Task 20: OffensEval 2020 (Zampieri et al. 2020) (multilingual data, covering English, Turkish, Arabic, Danish, Greek). Refining the definition of what constitute offensive language in absolute terms is not a trivial task and, possibly, not even a useful path to follow. Being a highly subjective phenomenon, refinements should take into account the social, ethnographic, and community-oriented dimensions of where the interaction takes place and calibrate the offensiveness of a messages according to the values of a specific community of speech.

A complimentary and faster solution to monitor the impact of subjective interpretation, while avoiding strongly prescriptive annotation guidelines, is the introduction of a hierarchical annotation format that focuses on the distinction of *how* the targeted language phenomenon is realized (Waseem et al. 2017). Building on top of the definition of offensiveness provided by Zampieri et al. (2019a), it is possible to introduce an explicitness layer to distinguish whether the message is unambiguously offensive in its potential or not. If this strategy is applied to a parallel annotation by two or more (trained) annotators, it will be easier to distinguish potential bias (i.e, an over annotation of explicit messages) and identify controversial cases (i.e., whether a message is offensive or not). Such a strategy, for instance, has been adopted for the annotation of DALC (Caselli et al. 2021b, Ruitenbeek et al. 2022) for both abusive and offensive language.

A further feature of offensive language is that it represents the starting point to define and identify other socially unacceptable language phenomena such as abusive language and hate speech. For instance, a "targeted insult" when addressed to a social group (e.g., women, LGBTQIA+ people) can correspond to an instance of hate speech while when directed towards an individual may be an instance of abusive language. A similar behavior can be associated with "threats". This makes it possible to connect together different phenomena and focusing on their differences and interactions (Poletto et al. 2021).

Datasets and corpora for offensive language are available in more than 10 languages across a varied representation of language families.[3] In most of the cases, the annotation of offensive language is accompanied by the presence of other more fine-grained phenomena such as sexism or hate speech (targeted at a specific social group). The majority of these corpora adopts Twitter as reference Social Media platform. Clearly, the availability of APIs and Terms of Use that easily allow scholars to obtain and re-share the data[4] makes Twitter a privileged channel for these types of studies. At the same

[2] `https://mtsu.edu/first-amendment/article/392/jacobellis-v-ohio`

[3] For more details see `https://hatespeechdata.com`

[4] A big limitation of Twitter-based studies is that corpora tend to shrink over time since users have the right to permanently delete their messages or because their account are suspended. Nevertheless, by adopting appropriate anonymization procedures and a Data Management Plan, the sharing of the full text of a tweet - rather than just the IDs - is possible. We are currently working to release the DALC data in this format.

time, the limited amount of characters that Twitter allows for a message, makes the study of forms of socially unacceptable language very limited in their variety and complexity (Vidgen and Derczynski 2020). A further common property of most offensive language datasets is the distribution of the positive and negative examples. The overall amount of socially unacceptable language phenomena on Social Media is unknown. Estimates on Twitter position it between 1% and 3% of the total messages (Founta et al. 2018) while other venues, such as GAB or `TheRedPill` subreddit, have a denser distribution of these phenomena. Nevertheless, messages containing socially unacceptable language phenomena are a minority with respect to the overall user-generated inputs. In an attempt to maintain an ecological representation of the data distribution, corpora present an unbalanced distribution of the positive and the negative classes, with a tendency to under-represent the positive class to a third of the entire corpus size.

Systems for offensive language detection are based on supervised machine learning methods, either using monolingual or multi-lingual learning strategies. In recent years, there is an growing tendency to reduce this (and other tasks) to a "simple" fine-tuning of pre-trained language models (PTLMs) (Zampieri et al. 2019d, Kumar et al. 2020, Pelicon et al. 2021, Gaikwad et al. 2021, y Jorge Carrillo-de-Albornoz y Laura Plaza y Julio Gonzalo y Paolo Rosso y Miriam Comet y Trinidad Donoso 2021). Not surprisingly, the use of fine-tuning results in state-of-the-art performances, with averages macro-F1 scores across different languages above 0.80. However, when dissecting the results per class, we observe that all systems tend to under-perform with respect to the positive class, with F1 scores in the a range between 0.60 and 0.70. Quite interestingly, the use of simpler architectures - such as Support Vector Machines - are very strong baselines for these tasks.

As for Ducth, previous work is quite limited. Van Hee et al. (2015) investigate cyberbullying focusing on the context in which an aggressive or hurtful message is posted. Posts are retrieved from a Dutch question answering social medium, are annotated using a three-point scale, where 0 indicates no cyberbullying, 1 the presence of context and indications of cyberbullying, and 2 indicates explicit and serious threats or incitements to commit suicide. A different contribution, focusing on racism, has been presented by Tulkens et al. (2016). In this work, the authors develop a dictionary-based system,HADES (HAte speech DEtection System), to detect racist messages in Dutch. The dictionary is created in a semi-automatic way in three steps starting with seed terms and then expanding them using a word2vec model. The last step is a manual curation step where unsuitable expansions are removed. Lastly, more recent work has been conducted in the area of offensive and abusive language thanks to the development of DALC.[5] The best system for offensive language developed on DALC with a PTLMs (RobBERT, Delobelle et al. (2020)) achieves a macro F1 score of 0.828, with an F1 on the positive class of 0.746 (Ruitenbeek et al. 2022).

2.2 Cross-lingual Transfer and Low-data Settings

Cross-lingual transfer learning is a relatively new task in NLP which has sparked lot of interest in the community (Schuster et al. 2019, Do and Gaspers 2019, Li et al. 2020, among others). The key idea is that dense representations (i.e., embeddings) of tokens or sentences are built into a common representation space, either using parallel data or not, that allow to train a system in one language on a specific task and transfer the learned model to other languages not present during the training phase. The benefit of the use of this method has been consistently shown in different tasks, from morpho-syntax to Machine Translation (Lynn et al. 2014, Aufrant et al. 2016, Ahmad et al. 2019, Kim et al. 2019, Chen et al. 2021). Recent work has advanced the state of the art by proposing innovative approaches such as adapter modules (Pfeiffer et al. 2020, Üstün et al. 2020) or the use of meta-learning (Lee et al. 2022). One of the proved advantages of the use of cross-lingual transfer learning is that it allows to apply NLP tools to less-resourced languages, even if never seen in training (i.e., zero-shot setting) or included in the development of the shared representations, obtaining satisfying performances. It would be an overstatement to claim that these models offer

[5] More details on the corpus are presented in Section 3

a definitive solution for languages that are under-resourced. First, there is a growing evidence showing how monolingual PTLMs outperform multilingual ones in every tasks and setting (de Vries et al. 2020). Secondly, performances of zero-shot experiments are actually boosted when adding few manually annotated data for the target language, even if the target language is not present in the pre-training materials (Zhou et al. 2021, Üstün et al. 2022). Finally, the data (i.e., size and quality) used to generate multilingual PTMLs have a relevant impact on the models' predictions in zero-shot settings due to differences in writing systems can degrade performance as well as cultural differences primarily encoded in natural language (Pires et al. 2019, Lauscher et al. 2020).

A complementary approach to zero-shot transfer is self-training (Blum and Mitchell 1998, McClosky et al. 2006). Self-training is a simple method to reduce data sparseness, especially in low data settings. Withing the self-training paradigm, an existing model is applied to a large set of unlabelled data. The newly labeled data are considered as ground-truth and recombined with the manually annotated one used for training (gold data) to develop a new model. Although the effectiveness of self-training is debatable with varying results, on the basis of the task and amount of data (UzZaman et al. 2013, Basile and Caselli 2020, Mi et al. 2021), it is interesting to investigate its use in the light of the advancements given by PTLMs and for a task such as offensive language detection.

3. Data

For this work, different datasets have been used and developed for training and testing our models. In the following sections we highlight the properties of the various datasets and discuss how the unlabelled data for Dutch have been collected.

3.1 Gold Training Data

We have collected two manually annotated datasets to train our models. The first is the Political Speech Project dataset (Bröckling et al. 2018) and the second is DALC. While the Political Speech Project dataset will represent the starting point for our experiments using transfer learning and self-training data, DALC is mainly used to develop an upper-bound threshold on the newly developed test set.

Political Speech Project The Political Speech Project (PSP, henceforth) dataset has been developed by a consortium of journalists in Europe whose primarily focus was to investigate abuse against female politicians and whether politicians from particular parties were more subject to abuse than others. The authors randomly sampled 320 politicians (40 men and 40 women) from four countries (France, Germany, Italy and Switzerland) using lists of parliamentary and government ministers. In addition, each journalist selected 10 prominent politicians (5 men and 5 women) from each country representing the full political spectrum. The data were collected during a 4-week period (February 21 – March 21, 2018) from Twitter and Facebook. A random sample of 41,222 posts has been manually annotated for two dimensions: (i) the level of offensiveness, and (ii) the category of offence that was expressed. The level of offensiveness is expressed via a scale of three values having at the extreme sides the values "neutral" - for non-offensive messages - and "highly offensive". The category of the offence is used to better identify the content of the offensive message. The authors adopt seven classes, including sexist, anti-immigrant, anti-Muslim, anti-Semitic, homophobic and other. The category labels are not mutually exclusive since offensive expressions may be intrinsically intersectional (Crenshaw 2017) targeting multiple dimensions at the same time (e.g., sexism and anti-Muslim , if the offence is addressed to a Muslim woman politician, for instance). In addition to this, a further annotation layer distinguishes whether the offence is personally directed to the politician mentioned in the message or not. Unfortunately, the accompanying documentation of the dataset does not contain definitions and guidelines for the use of the level category, leaving the interpretation to the annotators. Although there can be variability in the labels across single annotators, the dataset is uniform and spans across multiple languages representing a unique in the language resource panorama. Table 1 illustrates some examples from the PSP data.

Country	Text	Off. Level	Off. Category	Is personal?
Switzerland	Jacqueline Badran Nationalrätin SP (ZH) sagt #NEINzuNoBillag! @JayBadran https://t.co/6ZGu5Q1y8l	0	None	No
France	@flo_agri @auroreberge Et vous la gauchiasse avec plein de sang sur les mains et de nombreux meurtres de masse.	1	None	Other
Italy	@ckyenge @mattiellodavide Stai zitta st** di m**. .torna nella tuo c** di Congo	2	Sexist, Anti-immigrant, Other	Other
Germany	nunja ist ne ditib einrichtung das müssen nicht zwangsläufig afd leute gewesen sein am liebsten würd ich die einrichtungen auch abfackeln	3	Anti-immigrant, Anti-muslim	Other

Table 1: Examples from the PSP dataset, showing the country, the text of the message, the level of offensiveness (Off. Level), the category of the offence (Off. Category), and whether the offence directly targets the politicians or not (Is personal?)

Having adopted a random sampling approach for the annotations with no correction strategies, the PSP dataset presents a very skewed distribution of the labels across all annotated dimensions. In particular, the messages belonging to the negative class make up for 93% (38,516) of all the occurrences. A full overview of the distribution of the level of offensiveness per country is represented in Table 2.

Country	**Off. Level**				**Total**
	0	**1**	**2**	**3**	
France (FR)	12,401	589	225	48	13,263
Italy (IT)	12,297	777	151	19	13,244
Germany (DE)	9,296	425	106	18	9,845
Switzerland (CH)	4,522	258	79	11	4,870

Table 2: PSP data: Distribution of messages per rating and country (0: neutral, 1: mildly offensive, 2: offensive, 3: highly offensive).

As Table 2 shows, the skeweness also affects the distribution of messages per country, with France and Italy representing almost 64% of the total, while Germany provides 24% and Switzerland a mere 12%. This has an impact in the language distribution. Quite surprisingly, Switzerland has only messages in Swiss German. Although Swiss German is very closely related to Standard German, they are not the same. Variations in spelling, morpho-syntax, and vocabulary may all have an impact in the development of our transfer learning model. In addition to this, there are instances of code-switching. Using Langid[6] we found that the data for France and Germany can be multilingual, due to the presence of English phases or clauses.

To run our experiments, we converted the 4-way classification of the offence level to a binary format, by collapsing into a global OFFENSIVE category the labels 1, 2 and 3, and using a NOT class for the label 0. In addition to this, we excluded 3,499 messages because the full text was not properly retrieved or missing, leaving us with 37,723 entries. We present in Table 3 an overview of the re-arranged classes per country.

[6] `https://github.com/saffsd/langid.py`

Class	**Country**				**Total**
	FR	**IT**	**DE**	**CH**	
OFFENSIVE	274	164	233	138	809
NOT	11,184	11,793	9,352	4,585	36,914

Table 3: PSP data aggregated: distribution binary classes per country. Country names are expressed using ISO 3166 code.

The removal of messages lacking the full text had an impact on the positive class: offensive comments now account for only 1% of data. The limited number of positive examples poses a challenge to the development of our transfer learning model, requiring the application of udersampling with respect to the negative class.

DALC DALC (Caselli et al. 2021a, Ruitenbeek et al. 2022)[7] is a Twitter-based corpus in Dutch annotated for abusive and offensive language phenomena. DALC inherits the definition of offensive language used by Zampieri et al. (2019b). Three collection methods (keywords extraction, message geolocation, and seed users) have been employed to extract the messages that compose DALC. This strategy combined with the sampling of the data at different moments in time has been adopted as a strategy to limit developers' bias. The annotation of offensive language has been conducted by applying a set of semi-prescriptive annotation guidelines by four annotators in parallel. Following the proposal by Waseem et al. (2017), DALC applies a multi-layer annotation approach distinguishing between the explicitness of the message and its target. The explicitness layer addresses the surface form of the message in its potential to be (perceived as) offensive and distinguishes three classes: (i.) EXPLICIT; (ii.) IMPLICIT; and (iii.) NOT. The NOT class is used for not offensive messages. On the other hand, the EXPLICIT class applies to messages with profanities or that contain a combination of words that unambiguously make the message offensive. IMPLICIT messages are more subtle, lacking any surface markers, and thus making the offence hidden.

The target layer, on the other hand, indicates towards *whom* the offence is directed. Four classes have been identified: (i.) INDIVIDUAL, for messages that are addressed to or target a specific person or individual (who could be named or not); (ii.) GROUP, for messages that target a group of people considered as a unity because of ethnicity, gender, political affiliation, religion, disabilities, or other common properties; (iii.) OTHER, for messages that target concepts, institutions and organisations, or non-living entities; and (iv.) NOT, for offensive messages without a target.

The parallel annotation of the data offers a more comprehensive and diversified view of what constitutes offensive language. In case a majority in the labels for each of the layer is not present, annotators were instructed to discuss the case and find an agreement. An interesting feature of the offensive language dimension in DALC is the availability of both aggregated and per annotator labels to foster future work on the relationship of subjectivity and annotation of natural language phenomena (Basile 2020, Leonardelli et al. 2021). Table 4 presents three examples of messages in DALC annotated on the explicitness and the target layers.

When it comes to the distribution of labels, DALC is unbalanced - as all socially unacceptable language datasets - but in a very different way when compared to PSP. In particular, DALC is composed by 11,292 messages and it maintains a 2/3 *vs.* 1/3 split of negative and positive classes. Furthermore, training, development, and test splits for DALC are fixed and follows a strict division of the messages whereby no topic or time period that is in the training data appears in the test set. We present the distribution of the data in Table 5, excluding the target layer. The OFFENSIVE label is obtained by collapsing together the EXPLICIT and the IMPLICIT labels.

Although explicit messages are the majority, the number of implicit cases is relatively high - representing 41.28% of all occurrences. A possible explanation can be found in the annotation guidelines:

[7] `https://github.com/tommasoc80/DALC`

Text	**Explicitness**	**Target**
Dat gebeurt in het park en veel jongeren bij elkaar	NOT	NOT
@USER Hier ben ik het mee eens. Journalistiek hoort "onafhankelijk" te zijn en dat is in NL al lang niet meer aan de orde.	IMPLICIT	OTHER
@USER Ze doen zo schijnheilig. Hoeveel b** worden er niet vermoord al dan niet door politieagenten? Dan demonstreren ze niet, deze personen zijn g**	EXPLICIT	GROUP

Table 4: DALC: examples of annotated data. User handlers have been anonymized.

Data Split	**OFFENSIVE**		**NOT**
	EXP.	**IMP.**	
Train	1,407	1,070	4,340
Dev	230	209	766
Test	584	283	2,403
Total	3,783		7,509

Table 5: DALC: label distribution. EXP. = EXPLICIT; IMP, = IMPLICIT.

being only semi-prescriptive offensive messages have been labelled as such either because they contained a profanity, or because they were *perceived* as such by the annotators.

3.2 Unlabelled Dutch Data

To validate the impact of zero-shot transfer from multilingual PTLM and the contribution of automatically annotated data (silver data, henceforth), we have collected a large quantity of tweets that contains a mention of a Dutch politician. We created a list of politician twitter accounts using information from the *Tweede Kamer*[8] (i.e., the Dutch House of Representatives) in September 2021. At that time, the *Tweede Kamer* consisted of 148 politicians from 18 different parties, with 4 parties having more than 10 seats. The full list is available on the DALC GitHub.

We then retrieved tweets containing mentions of politicians using the RUG Twitter Corpus (Sang 2011). In particular, we collected tweets from the Netherlands for the whole month of March 2021. This month is particularly interesting because it contains a relevant historical event such as the Dutch Parliamentary elections.[9] At that time, people in the Netherlands were still experiencing lockdowns and limitations of civil liberties to contrast the COVID-19 pandemic. Additionally, the vaccination campaign against COVID-19 was experiencing delays due to bottlenecks in dose distributions and organization. The presence of the elections and the measures to counteract the ongoing COVID-19 pandemic makes it a perfect combination to trigger many messages mentioning or targeting politicians.

To select only relevant messages, we identified four methods in which a user could mention a politician:

- direct use of the politician Twitter handler (`dm-tweet`); e.g., "@markrutte";
- mention the full name (first and last name) of a politician (`fn-tweet`); e.g., "Mark Rutte";
- mention of a politician's last name in an hashtag (`hln-tweet`); e.g. "#rutte";
- mention of the last name only of a politician (`ln-tweet`); e.g., "Rutte";

[8] `https://www.tweedekamer.nl/kamerleden_en_commissies/alle_kamerleden`

[9] The elections took place on 17 March, 2021.

The combination of these strategies returned a total of almost 2 million unique messages. Table 6 presents the number of the extracted tweets per method.

Method	**Retrieved tweets**
`dm-tweet`	859,676
`fn-tweet`	61,239
`hln-tweet`	71,277
`ln-tweet`	940,896
Total	1,933,088

Table 6: Distribution of the different tweet types that were collected

Not surprisingly, the `dm-tweet` strategy returns the largest amount of valid data, being the most precise way to identify messages directed at politicians. On the other hand, `ln-tweet` is the most noisy since the last name in a tweet can refer to someone else other than a politician. We thus decided to exclude all messages retrieved in this way.

The `fn-tweet` and the `hln-tweet` return in proportion less messages although being quite accurate in their results. Clearly, the fewer instances retrieved for `fn-tweet` may be due to the peculiarities of Twitter: having a limiting amount of characters to express oneself the use of full name of a politician is rather useless. At the same, transforming a politician last name in an hashtag may be either an attempt to push the visibility of the message rather than targeting a politician's attention, or a strategy to bypass a block on Twitter by the targeted politician. The final size of the unlabelled Twitter data amounts to 992,192.

4. Experiments

In this section, we will present the experiments we have run to further validate our transfer learning approach and the impact of silver data. As a by-product of running a fine-tuned multilingual model on the PSP dataset, we developed a new manually annotated benchmark for Dutch. While this new benchmark is still based on the same medium as DALC, it represents a more specialized language variety when compared to DALC since all messages contain at least the mention of one politician and cover a time period not present in DALC. For our experiments we used two PTLMs: mBERT (Pires et al. 2019)[10] and BERTje (de Vries et al. 2019).[11] In all experiments we have used the same basic pre-processing steps by means of regular expressions. Details are in Appendix A.

4.1 Zero-shot Transfer with PSP Data

As we have illustrated in Section 3.1, the PSP dataset is heavily imbalanced towards the negative class. A preliminary experiment where we fine-tuned mBERT (Pires et al. 2019) using the distribution of the data as is and applied 10-fold cross-evaluation resulted in the system not being able to learn any instances from the positive class. We thus applied under-sampling of the negative class and up-sampling of the positive one to balance the class distribution to boost the robustness of the system towards the positive class. Given the limited amount of data, we decided to create a challenging test set where positive and negative classes are balanced in a 50-50 proportion (236 messages for the negative class and 247 for the positive class). Keeping the 50-50 split, we ended up with a total of 3,413 messages, where the positive class corresponds to 1,706 messages and the negative class to 1,707.

[10] We used `bert-base-multilingual-cased` .

[11] BERTje is available at `GroNLP/bert-base-dutch-cased` .

With this peculiar training setting, we then tested whether learning of offensive messages is actually enhanced and to what extent. We compared the fine-tuning of mBERT[12] against a linear SVM.[13] The SVM is used as a baseline to compare the results of the mBERT model and as a lower threshold on the PSP data following our split. Results are reported in Table 7.

System	**Class**	**Precision**	**Recall**	**Macro F1**
SVM	NOT	0.716	0.737	0.728
	OFF	0.741	0.720	
mBERT	NOT	0.777	0.754	**0.774**
	OFF	0.771	0.793	

Table 7: Results on the PSP dataset with 50-50 proportion on test. Evaluation metrics are Precision and Recall per class. Macro-average F1 is used as global evaluation measure to rank the. systems. OFF = offensive; NOT = not offensive

The results indicate that the SVM is already a competitive baseline, although it under-performs on all classes when compared to mBERT. Quite interestingly, mBERT obtains very good results on the positive class, with a Recall score even higher than that it has for the non offensive messages. The relatively high results on this test set should also be taken as a potential warning of the presence of some bias in the data in terms of similarities and potential lexical overlap between the training and test materials. Nevertheless, the results of the experiments are positive. This allows us to move forward and apply the fine-tuned mBERT model on the PSP data (`mBERT-PSP`) to the unlabelled Dutch data with mentions of politicians.

The Offend the Politicians Benchmark (OP-NL) As we have previously anticipated, the development of the new dynamic benchmark for offensive language targeting politicians in Dutch is a by-product of the deployment of the `mBERT-PSP` model over the 900k tweets we collected. Once we deployed `mBERT-PSP` on the unlabelled Dutch tweets, we randomly extracted 1,500 messages and manually corrected them. We will refer to this new test data as the Offend the Politicians Benchmark, henceforth OP-NL.

The manual correction applied the definition of offensive language and the annotation guidelines used for the development of DALC. The annotation/correction was conducted by one annotator and it focused only on the distinction between offensive and not offensive, without applying the full hierarchical annotation strategy used in DALC.[14] Table 8 illustrates the results of the correction process by means of a 2x2 confusion matrix, where N stands for the non-offensive messages, and P for the offensive ones. As the Table shows, the predictions of `mBERT-PSP` are unbalanced, with a tendency to predict a message to be not offensive (1,023 *vs.* 477). The manual correction clearly confirms this, showing that the largest class that underwent correction of labels where messages wrongly predicted as not offensive (280 messages), while only a minority of non-offensive messages (218) was wrongly predicted as offensive. After correction, the distribution of labels is as follows: 961 messages (64%) are not offensive and 539 (36%) are offensive. The ratio between non-offensive and offensive messages is 1.78 : 1, very close to the label distribution in DALC.

In terms of annotation efforts, the predictions have to be corrected only for 498 cases (218 + 280 misclassifications) out of the total of 1,500 messaged, representing less than $\frac{1}{3}$ of all occurrences. This has significantly sped up the development of OP-NL.

[12] We used the default parameters for learning rate and optimizer from the `simpletransformers` library, batch size = 16, number of training epochs = 5 .

[13] We have used the `scikit-learn` library for the SVM, with default value for the C parameter, TF-IDF vectorization, with unigrams and bigrams for tokens, and trigrams and fivegrams for charatcters.

[14] The annotator is one of the authors of this paper. Full data statement for OP-NL is reported in the DALC repository.

		Prediction		
		N	**P**	**Total Human**
Manual correction	**N**	**743**	218	961
	P	280	**259**	539
Total `mBERT-PSP`		1,023	477	1,500

Table 8: OP-NL: Confusion matrix of the manual correction with respect to `mBERT-PSP` predictions.

To validate the annotation, we performed an inter-annotator agreement (IAA) study with a second annotator on a subset of 150 messages (10%). This resulted in a Cohen's kappa of $\kappa = 0.67$, indicating substantial agreement and in line with previous work. Although the IAA score is quite satisfying, the whole of OP-NL is annotated by only one person. Since offensive language is a highly subjective phenomenon, OP-NL has been enriched with a further annotation layer expressing the annotator's confidence on the assigned label. Confidence levels are expressed by means of a Likert-scale from 1 (very uncertain) to 4 (very certain). The confidence levels are distributed as shown in Table 9. Most labels (83.6%) are given high certainty values (3 or 4), and only in 5% of the cases the annotator was highly uncertain (1) about the assigned label. Although this does not fully address the subjective interpretation of (some) of the messages, we consider the presence of the confidence scores as a strategy to address this issue and of help for future work and extension of the dataset.

Level 1	**Level 2**	**Level 3**	**Level 4**	**Total**
84	158	160	1,094	1,500

Table 9: OP-NL: Distribution of confidence levels of labels.

Evaluating `mBERT-PSP` against OP-NL Having created a reference gold standard for Dutch, we can now evaluate the predictions in a zero-shot setting of `mBERT-PSP`. Results are illustrated in Table 10.

System	**Class**	**Precision**	**Recall**	**Macro F1**
`mBERT-PSP`	NOT	0.726	0.773	0.629
	OFF	0.543	0.480	

Table 10: Results of `mBERT-PSP` on the gold version of OP-NL.

Overall, `mBERT-PSP` obtains good performances, especially on the non offensive messages, even if Dutch has never been seen at training time and considering that the data composing PSP comes from two Social Media platforms, namely Facebook and Twitter, with different properties in terms of how a message can be posted (e.g., length, presence of hashtags, mentions of users, among others). On the other hand, we observe that the model struggles with predicting the offensive class, with a

F1 score for this class barely above 50% (0.509). In the reminder of our experiments, our goal will be to boost the performance on the OP-NL dataset by means of silver data in Dutch.

4.2 Impact of monolingual silver data with mBERT

It is known that having, at training time, data from the language (and task) in which the model will be tested is helpful. The following experiments investigate whether adding monolingual silver data to a manually annotated dataset such as PSP confirms this effect or not. In addition to this, it is also relevant to understand whether a potential beneficial effect of silver data is dependent on the amounts that are added to the gold data.

Our first experiment is partially inspired by Ranasinghe and Zampieri (2020) and it adopts an incremental self-training approach to expand the size of the silver data and test their impact. We call this method `dyna-mBERT-PSP`. The idea is quite simple: over 10 iterations, a total 3,410 silver data instances, corresponding to almost the same size of the PSP training, are used to fine-tune `mBERT-PSP`. At each iteration, 10% of the total data (i.e., 341 instances) are used to fine-tune the model. The dynamic part of the method is that a fine-tuning iteration represents a predict-train-predict cycle where, firstly, the weights of an initial fine-tuned model, M_t, are used to predicts the labels of 341 data instances in Dutch. Secondly, the weights of the model are updated by fine-tuning the additional labels of the predicted 341 instances, creating a new model M_{t+1}. Finally, the newly fine-tuned model, M_{t+1}, is evaluated against OP-NL. After this, an new cycle is started to predict labels for additional silver training data. The cycle stops after 10 iterations. In all iterations, we maintained the seed fixed.

The second method we implemented is slightly different. We call this `mBERT-PSP+`. The method is still based on self-training, but in this case, rather than saving the weights of the multilingual model (i.e., `mBERT-PSP`) and then fine-tune it on target language silver data, we initially add a total of 3,413 automatically labeled messages in Dutch to the original PSP data. The initial training set is thus composed by 6,826 messages. We then used this training corpus, a hybrid of gold and silver data, to fine tune a new model and evaluate it against OP-NL. The main difference of this approach with respect to `dyna-mBERT-PSP` consists in implementing a train-predict cycles only. After this initial "kick start" with a larger set of silver data, we keep expanding the amount of Dutch silver data in training by blocks of 1,000 message at each iteration. Being a train-predict cycle only, at each iteration a new model is trained with an increasing amount of training material. At iteration 10, `mBERT-PSP+` is trained on a combination of 3,413 messages from the PSP dataset and 13,413 Dutch silver data. The rationale for this set of experiments is to obtain more evidence of the impact of silver data. Like for the `dyna-mBERT-PSP` experiments, the cycles stops after 10 iterations. In all iterations, we maintained the seed fixed. In both experiments, there is no overlap in the selected silver data across all iterations. The results for both experiments are presented in Table 11.

In general, we observe a degrading of the performances in both experiments when silver data are used. However, there are some peculiarities and differences. For `dyna-mBERT-PSP`, in all experiments, we note that silver data play a positive role in increasing the Recall for the negative class and, to a smaller extent, the Precision for the positive one. The changes in the macro-F1 have a random behaviour at the beginning and until iteration 5, and they consistently degrade in the remaining iterations. However, it seems that at iteration 4 we reach a nice spot obtaining a macro-F1 score of 0.634 with a delta of 0.005 points in favor `dyna-mBERT-PSP` when compared to `mBERT-PSP` (see iteration 0).

When looking at the results for `mBERT-PSP+`, the impact of silver data seems to be less drastic. Although we add 1,000 new messages at each iteration, the macro-F1 score remains quite stable in its degradation, with a less random behavior than `dyna-mBERT-PSP`. When compared to `dyna-mBERT-PSP`, silver data, in this case, tend to increase the Precision with a limited impact on the Recall scores for the positive class. At the same time, the impact on the negative class seems more varied. Until iteration 5, Recall increases at the expenses of Precision. This drastically changes

		dyna-mBERT-PSP			mBERT-PSP+		
Training iterations	**Class**	**Prec.**	**Rec.**	**Macro F1**	**Prec.**	**Rec.**	**Macro F1**
0	NOT	0.726	0.773	0.629	0.719	0.783	0.622
	OFF	0.543	0.480		0.541	0.456	
1	NOT	0.721	0.796	0.627	0.714	0.826	0.622
	OFF	0.554	0.452		0.570	0.411	
2	NOT	0.695	0.860	0.592	0.714	0.832	0.623
	OFF	0.569	0.328		0.577	0.408	
3	NOT	0.696	0.860	0.595	0.714	0.832	0.623
	OFF	0.571	0.332		0.577	0.408	
4	NOT	0.725	0.806	**0.634**	0.712	0.805	0.615
	OFF	0.568	0.454		0.547	0.419	
5	NOT	0.711	0.808	0.615	0.717	0.807	0.624
	OFF	0.549	0.415		0.558	0.434	
6	NOT	0.696	0.840	0.593	0.723	0.757	0.627
	OFF	0.550	0.346		0.533	0.493	
7	NOT	0.694	0.856	0.590	0.725	0.784	**0.631**
	OFF	0.561	0.328		0.551	0.471	
8	NOT	0.693	0.887	0.589	0.719	0.774	0.621
	OFF	0.600	0.300		0.543	0.462	
9	NOT	0.691	0.903	0.585	0.721	0.783	0.625
	OFF	0.620	0.282		0.543	0.460	
10	NOT	0.683	0.912	0.566	0.717	0.783	0.619
	OFF	0.612	0.246		0.537	0.449	

Table 11: Results for the `dyna-mBERT-PSP` and `mBERT-PSP+` over the full 10 iterations on OP-NL. Macro-average F1 is used to identify the best performing models. Precision and Recall per class are reported as well. Best results for each model are marked in bold.

from iteration 6 onward where we observe a drop in Recall but limited losses for the Precision. It is as if the old saying "the more data, the better" seems to maintain some of its efficacy in this setting. At iteration 7, with 7,000 extra silver data in the training, we reach a macro-F1 of 0.631, beating the regular `mBERT-PSP`, but still below `dyna-mBERT-PSP`.

The fact that in both settings we observe an improvement of the macro F1 score at a certain point suggests the possibility of identifying a threshold in terms of silver data material that could be added to an initial model to boost its performances in a positive direction. This, unfortunately, is more a working hypothesis than a confirmed findings from our experiments.

4.3 Using BERTje: Impact of gold and silver data

So far, all experiments we have conducted make use of mBERT, a multilingual PTLMs. Limitations in performance of multilingual PTLMs when compared to corresponding monolingual models are known (Rust et al. 2021). Given that Dutch has available monolingual PTLMs,[15] we decided to investigate the use of silver data generated by `mBERT-PSP` when used to train a monolingual PTLM.

[15] At the time of writing, available PTLMs for Dutch are BERTje (de Vries et al. 2019), RobBERT (Delobelle et al. 2020), and RobBERTje (Delobelle et al. 2021).

For our experiments, we selected BERTje (de Vries et al. 2019) because of its similarity, in terms of architecture, with mBERT and the careful data selection process followed for its development.[16]

BERTje allows us to run multiple experiments with different data combinations. Firstly, we can assess a potential upper-bound on OP-NL when gold data are used to fine-tune a model. Secondly, we can asses what is the performance of fine-tuning a model using silver data only. Finally, we can compare the results of these models when the training set is composed by silver and gold data.

Upper-bound on OP-NL The OP-NL benchmark is not accompanied by a corresponding training corpus in the spirit of dynamic benchmarking. To assess a potential upper-bound performance on this benchmark, we have used the training data from DALC to fine-tune BERTje. The advantage of this model with respect to all the others is that it uses monolingual data annotated for the same language phenomenon of the OP-NL benchmark. Furthermore, the definition of what constitutes an offensive message in DALC is the same we have applied to OP-NL. The challenge is that the data come from a different time period and topic than those used to create the training set of DALC. We refer to this model as `BERTje-Gold`.

Silver data only In this setting, BERTje is fine-tuned with silver data originally predicted with `mBERT-PSP`. In particular, we selected 3,143 silver messages replicating the size (but not the label distribution) of the PSP data. This model is called `BERTje-Silver` In addition to this, we replicated the dynamic self-training setting of `dyna-mBERT-PSP`. In this case, for each predict-train-predict cycle we use 1,000 silver labeled messages. The iterations stop after 10 cycles. It is important to stress that in this case, with exclusion of the initial set of 3,143 silver messages (iteration 0), the labels for the messages of each new iterations (iterations 1 to 10) are obtained using BERTje. We call the models obtained with this approach `dyna-BERTje-Silver`. In all iterations with `dyna-BERTje-Silver`, we did not change the seed.

Extend gold with silver The last experiment consists in expanding the gold data from DALC with additional silver data. We have followed two settings by differentiating the size of the silver data. In the first case, we added the same amount of silver messages as the size of the PSP training data, i.e., 3,413 messages. The resulting model is called `BERTje-Silver-3k`. In the second variation, we used an amount of silver data that was the double of the PSP training, i.e., 6,826 messages, reaching an amount of data comparable to the manually annotated training in DALC (i.e, 6,817). We will refer to this model as `BERTje-Silver-6k`. In both cases, the silver data are obtained from the predictions of `mBERT-PSP`. For each of these experiments, the training data have been shuffled before passing them to BERTje to avoid any impact related to the order in processing the gold and silver data during the fine-tuning.

We present an overview of the results of all the experiments in Table 12. For reason of space and to maintain a coherent comparison of the results across the different experiments, we report the best results for the `dyna-BERTje-Silver`. The complete overview of the `dyna-BERTje-Silver` for each iteration is presented in Appendix B.

Not surprisingly, the use of gold data leads to the best results across all experiments with BERTje, reaching a macro-F1 of 0.737, with very balanced scores for Precision and Recall for the positive class (Precision = 0.677 and Recall = 0.641). In general, the results of models fine-tuned on silver data are quite disappointing if interpreted only in terms of performance since no system obtains scores that are competitive with `BERTje-Gold`. On the other hand, when considering `BERTje-Gold` as an ideal upper-bound limit of potential performance on OP-NL, the picture is a quite different. First, we observe a positive effect of the `dyna-BERTje-Silver` approach when compared to `BERTje-Silver`, indicating a positive effect of the dynamic self-training approach within a monolingual PTLM.

[16] For all fine-tuning experiments with BERTje, we used the default parameters for learning rate and optimizer from the `simpletransformers` library, batch size = 8, number of training epochs = 5 .

System	Class	Precision	Recall	Macro F1
BERTje-Gold	NOT	0.804	0.828	**0.737**
	OFF	0.677	0.641	
BERTje-Silver	NOT	0.736	0.775	0.643
	OFF	0.558	0.506	
dyna-BERTje-Silver	NOT	0.749	0.809	0.667
	OFF	0.603	0.517	
BERTje-Silver-3k	NOT	0.768	0.813	0.691
	OFF	0.628	0.562	
BERTje-Silver-6k	NOT	0.761	0.816	0.684
	OFF	0.624	0.543	

Table 12: Results of the experiments fine-tuning using BERTje with gold and/or silver data. The results for dyna-BERTje-Silver correspond to training iteration #3. Best results are in bold.

Among all systems using silver data, the most competitive are BERTje-Silver-3k and BERTje-Silver-6k. Unfortunately, their results are different from what expected. The working hypothesis was based on a positive effect of the silver data material when joined to the gold data material, more or less in line with the effects seen with mBERT-PSP+. The contrary is actually the case. The best macro F1 is obtained by BERTje-Silver-3k, where the silver data are roughly half of the gold material. Using an amount of silver data which is almost the size of the gold training is actually detrimental. The macro F1 for BERTje-Silver-6k is lower by 0.007 points when compared to BERTje-Silver-3k.

5. Discussion

The picture that emerges from the experiments we have presented so far is actually more complex than it can appear at first sight. To facilitate the summary of our findings, we report in Table 13 the best results of all the models. In this overview table, BERTje-Gold serves as reference of a potential upper bound limit on OP-NL.

One of the substantial conclusions we can draw from the overview of the results is that silver data, although limited, do have a positive contribution for developing systems for language phenomena for which (eventually) no gold data is available. This is in line with previous findings (Socha 2020). The positive impact of silver material is amplified if directly used within a monolingual model rather in a multilingual one. This finding provides further evidence of the superiority of monolingual language models over multilingual ones.

Dynamic self-training appears to be a viable strategy to improve system performance. Besides observing the same trend with respect to the advantages of using a monolingual model with respect to a multilingual one, a matter that needs further investigation is the optimal amount of silver data to be used. On the basis of our experiments, there is an consistent trend that using an amount of silver messages in the range between 2,500 and 3,000 instances returns optimal results.

On the other hand, mixing silver data with gold data is detrimental. The results of BERTje-Silver-3k and BERTje-Silver-6k are pretty clear in this sense. Nevertheless, a point of reflection that we want to stress concerns the quality of the material produced by mBERT-PSP. The manual correction we conducted to create OP-NL indicates that a third of the predictions are incorrect. It could be the case that different results for the inclusion of silver data with gold data could be observed, if systems with better performances were used.

Finally, zero-shot transfer learning qualifies as a very useful method to boost the development of new gold data material. We have estimated that the creation of OP-NL took at least one third less of the time that it would have required, were we to annotate all data from scratch.

Training data	Model	Class	Precision	Recall	Macro F1
Silver	mBERT-PSP	NOT	0.726	0.773	0.629
		OFF	0.543	0.480	
	dyna-mBERT-PSP	NOT	0.725	0.806	0.634
		OFF	0.568	0.454	
	mBERT-PSP+	NOT	0.725	0.784	0.631
		OFF	0.551	0.471	
	BERTje-Silver	NOT	0.736	0.775	0.643
		OFF	0.558	0.506	
	dyna-BERTje-Silver	NOT	0.749	0.809	0.667
		OFF	0.603	0.517	
Silver & Gold	BERTje-Silver-3k	NOT	0.768	0.813	0.691
		OFF	0.628	0.562	
	BERTje-Silver-6k	NOT	0.761	0.816	0.684
		OFF	0.624	0.543	
Gold	BERTje-Gold	NOT	0.804	0.828	0.737
		OFF	0.677	0.641	

Table 13: Summary of results of all models on the OP-NL benchmark. The results for dyna-mBERT-PSP correspond to training iteration #4. The results for mBERT-PSP+ correspond to training iteration #7. The results for dyna-BERTje-Silver correspond to training iteration #3.

6. Conclusion and Future Work

This work presents a thorough investigation on the use of silver data in different transfer learning settings, ranging from pure zero-shot cases to extending manually annotated data for training. In addition to this, we present and make publicly available a new benchmark, OP-NL. OP-NL is composed by 1,500 messages from Twitter annotated for offensive language and containing at least one mention of a Dutch politician.

More in details, we have used a multilingual but domain specific dataset, PSP, to train a multilingual language model (mBERT) to generate silver data. Different strategies on how to use the generated data have been implemented, including direct re-use of the silver data as is and dynamic self-training. We have further compared the impact of silver data when using a multilingual language model and a monolingual one. The results we have obtained clearly indicates that silver data have a more positive impact in system performance when used to fine-tune a monolingual model rather than when applied to a multilingual one. In addition to this, a large amount of silver data used to increase the size of training material does not always result in better performances. This is confirmed by the results of models developed using the dynamic self-training approach. In both cases, it appears that an optimal amount of new training material is around 3,000 instances, whereas using larger quantity of data degrades the performances. The quantity of training material used in each prediction-training-prediction cycle has also an impact on the number of iteration needed to obtain optimal results. Currently, our results are based on empirical observations, but further investigation is needed to validate these findings and potentially be able to predict them on the basis of the task and the quality of the silver data. Dynamic self-training has also an advantage over the addition of silver data in successive steps, as indicated by the results for mBERT-PSP+ in Table 11.

A clear indication of the negative effects of mixing silver and gold data comes from the experiments with BERTje-Silver-3k and BERTje-Silver-6k. In this case, regardless of the size of the silver data, the impact is detrimental.

On the other hand, zero-shot transfer learning has a beneficial effect in speeding the annotation process for developing new language specific dataset/corpora. As a result of the application of zero-shot transfer, we were able to create a new benchmark for offensive language Dutch, OP-NL. OP-NL will provide an additional benchmark to evaluate systems for offensive language in Dutch by maintaining the same medium of the training data (i.e., Twitter), but varying topics and time period. The resulting performances on OP-NL could be used to better spot weaknesses of systems in terms of robustness and portability. The OP-NL test set will be released as part of DALC.

Future directions will investigate the creation of high quality silver data. This can help both as a strategy to boost the development of benchmarks for language phenomena not yet available in a target language, and to develop robust baseline models. A direction to explore could be the use of multilingual language models fine-tuned with larger and good quality datasets that present detailed and specific definitions of the target language phenomenon. For the specific case of offensive language, a promising starting point are the corpora released for the SemEval 2020 Task 12: SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (Zampieri et al. 2020).

References

Ahmad, Wasi, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng (2019), On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing, *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, pp. 2440–2452. https://aclanthology.org/N19-1253.

Artetxe, Mikel and Holger Schwenk (2019), Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, *Transactions of the Association for Computational Linguistics* **7**, pp. 597–610, MIT Press, Cambridge, MA. https://aclanthology.org/Q19-1038.

Aufrant, Lauriane, Guillaume Wisniewski, and François Yvon (2016), Zero-resource dependency parsing: Boosting delexicalized cross-lingual transfer with linguistic knowledge, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, The COLING 2016 Organizing Committee, Osaka, Japan, pp. 119–130. https://aclanthology.org/C16-1012.

Basile, Angelo and Tommaso Caselli (2020), Protest event detection: When task-specific models outperform an event-driven method, *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, pp. 97–111.

Basile, Valerio (2020), It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks, *DP@AI*IA*.

Blum, Avrim and Tom Mitchell (1998), Combining labeled and unlabeled data with co-training, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, Association for Computing Machinery, New York, NY, USA, p. 92–100. https://doi.org/10.1145/279943.279962.

Bose, Tulika, Irina Illina, and Dominique Fohr (2021), Unsupervised domain adaptation in cross-corpora abusive language detection, *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics, Online, pp. 113–122. https://aclanthology.org/2021.socialnlp-1.10.

Bröckling, Marie, Vincent Coquaz, Alexander Fanta, Alison Langley, Mauro Munafò, Julian Pütz, Francesca Sironi, Leo Thüer, and Rania Wazi (2018), Political speech project. https://rania.shinyapps.io/PoliticalSpeechProject/.

Caselli, Tommaso, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim (2021a), DALC: the Dutch abusive language corpus, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Association for Computational Linguistics, Online, pp. 54–66. https://aclanthology.org/2021.woah-1.6.

Caselli, Tommaso, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim (2021b), Dalc: the dutch abusive language corpus, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 54–66.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert (2017), You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, *Proceedings of the ACM on Human-Computer Interaction* **1** (CSCW), pp. 1–22, ACM New York, NY, USA.

Chen, Guanhua, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei (2021), Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 15–26. https://aclanthology.org/2021.emnlp-main.2.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini (2021), The echo chamber effect on social media, *Proceedings of the National Academy of Sciences* **118** (9), pp. e2023301118, National Acad Sciences.

Crenshaw, Kimberlé W (2017), *On intersectionality: Essential writings*, The New Press.

de Vries, Wietse, Andreas van Cranenburgh, and Malvina Nissim (2020), What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, pp. 4339–4350. https://aclanthology.org/2020.findings-emnlp.389.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim (2019), Bertje: A dutch bert model, *arXiv preprint arXiv:1912.09582.*

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020), Robbert: a dutch roberta-based language model, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3255–3265.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2021), Robbertje: A distilled dutch bert model, *Computational Linguistics in the Netherlands Journal* **11**, pp. 125–140.

Dhanya, LK and Kannan Balakrishnan (2021), Hate speech detection in asian languages: A survey, *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, Vol. 1, IEEE, pp. 1–5.

Do, Quynh and Judith Gaspers (2019), Cross-lingual transfer learning with data selection for large-scale spoken language understanding, *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, pp. 1455–1460. https://aclanthology.org/D19-1153.

Founta, Antigoni Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis (2018), Large scale crowdsourcing and characterization of twitter abusive behavior, *Twelfth International AAAI Conference on Web and Social Media.*

Gaikwad, Saurabh Sampatrao, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan (2021), Cross-lingual offensive language identification for low resource languages: The case of Marathi, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, INCOMA Ltd., Held Online, pp. 437–443. https://aclanthology.org/2021.ranlp-1.50.

Haidt, Jonathan (2022), Why the past 10 years of american life have been uniquely stupid, *The Atlantic*.

Huang, Kuan-Hao, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang (2021), Improving zero-shot cross-lingual transfer learning via robust training, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 1684–1697. https://aclanthology.org/2021.emnlp-main.126.

Karouzos, Constantinos, Georgios Paraskevopoulos, and Alexandros Potamianos (2021), UDALM: Unsupervised domain adaptation through language modeling, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 2579–2590. https://aclanthology.org/2021.naacl-main.203.

Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams (2021), Dynabench: Rethinking benchmarking in NLP, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 4110–4124. https://aclanthology.org/2021.naacl-main.324.

Kim, Yunsu, Yingbo Gao, and Hermann Ney (2019), Effective cross-lingual transfer of neural machine translation models without shared vocabularies, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 1246–1257. https://aclanthology.org/P19-1120.

Kumar, Ritesh, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri (2020), Evaluating aggression identification in social media, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, European Language Resources Association (ELRA), Marseille, France, pp. 1–5. https://aclanthology.org/2020.trac-1.1.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš (2020), From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 4483–4499. https://aclanthology.org/2020.emnlp-main.363.

Lee, Hung-yi, Shang-Wen Li, and Thang Vu (2022), Meta learning for natural language processing: A survey, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Seattle, United States, pp. 666–684. https://aclanthology.org/2022.naacl-main.49.

Leonardelli, Elisa, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli (2021), Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 10528–10539. https://aclanthology.org/2021.emnlp-main.822.

Li, Zheng, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang (2020), Learn to cross-lingual transfer with meta graph learning across heterogeneous languages, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 2290–2301. https://aclanthology.org/2020.emnlp-main.179.

Lynn, Teresa, Jennifer Foster, Mark Dras, and Lamia Tounsi (2014), Cross-lingual transfer parsing for low-resourced languages: An Irish case study, *Proceedings of the First Celtic Language Technology Workshop*, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp. 41–49. https://aclanthology.org/W14-4606.

McClosky, David, Eugene Charniak, and Mark Johnson (2006), Effective self-training for parsing, *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, Association for Computational Linguistics, pp. 152–159. https://dl.acm.org/doi/pdf/10.3115/1220835.1220855.

Mi, Fei, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings (2021), Self-training improves pre-training for few-shot learning in task-oriented dialog systems, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 1887–1898. https://aclanthology.org/2021.emnlp-main.142.

Nooralahzadeh, Farhad, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein (2020), Zero-shot cross-lingual transfer with meta learning, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 4547–4562. https://aclanthology.org/2020.emnlp-main.368.

Pelicon, Andraž, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak (2021), Zero-shot cross-lingual content filtering: Offensive language and hate speech detection, *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, Association for Computational Linguistics, Online, pp. 30–34. https://aclanthology.org/2021.hackashop-1.5.

Pfeiffer, Jonas, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych (2020), AdapterHub: A framework for adapting transformers, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, pp. 46–54. https://aclanthology.org/2020.emnlp-demos.7.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019), How multilingual is multilingual bert?, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.

Poletto, Fabio, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco (2017), Hate speech annotation: Analysis of an italian twitter corpus, *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, Vol. 2006, CEUR-WS, pp. 1–6.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti (2021), Resources and benchmark corpora for hate speech detection: a systematic review, *Language Resources and Evaluation* **55** (2), pp. 477–523, Springer.

Ramponi, Alan and Barbara Plank (2020), Neural unsupervised domain adaptation in NLP—A survey, *Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 6838–6855. https://aclanthology.org/2020.coling-main.603.

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

Ruitenbeek, Ward, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli (2022), “zo grof !”: A comprehensive corpus for offensive and abusive language in Dutch, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Association for Computational Linguistics, Seattle, Washington (Hybrid), pp. 40–56. https://aclanthology.org/2022.woah-1.5.

Rust, Phillip, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych (2021), How good is your tokenizer? on the monolingual performance of multilingual language models, *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)*, Association for Computational Linguistics, Online, pp. 3118–3135. https://aclanthology.org/2021.acl-long.243.

Sang, Erik Tjong Kim (2011), Het gebruik van twitter voor taalkundig onderzoek, *TABU: Bulletin voor Taalwetenschap* **39** (1/2), pp. 62–72.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith (2019), The risk of racial bias in hate speech detection, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 1668–1678. https://aclanthology.org/P19-1163.

Schuster, Sebastian, Sonal Gupta, Rushin Shah, and Mike Lewis (2019), Cross-lingual transfer learning for multilingual task oriented dialog, *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, pp. 3795–3805. https://aclanthology.org/N19-1380.

Seering, Joseph, Tony Wang, Jina Yoon, and Geoff Kaufman (2019), Moderator engagement and community development in the age of algorithms, *New Media & Society* **21** (7), pp. 1417–1443, SAGE Publications Sage UK: London, England.

Socha, Kasper (2020), KS@LTH at SemEval-2020 task 12: Fine-tuning multi- and monolingual transformer models for offensive language detection, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), pp. 2045–2053. https://aclanthology.org/2020.semeval-1.270.

Thrush, Tristan, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela (2022), Dynatask: A framework for creating dynamic AI benchmark tasks, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Dublin, Ireland, pp. 174–181. https://aclanthology.org/2022.acl-demo.17.

Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans (2016), The automated detection of racist discourse in dutch social media, *Computational linguistics in the Netherlands journal* **6**, pp. 3–20.

Üstün, Ahmet, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord (2020), UDapter: Language adaptation for truly Universal Dependency parsing, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 2302–2315. https://aclanthology.org/2020.emnlp-main.180.

Üstün, Ahmet, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord (2022), UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling, *Computational Linguistics* **48** (3), pp. 555–592, MIT Press, Cambridge, MA. https://aclanthology.org/2022.cl-3.3.

UzZaman, Naushad, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky (2013), SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 1–9. https://aclanthology.org/S13-2001.

Van Hee, Cynthia, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste (2015), Detection and fine-grained classification of cyberbullying events, *Proceedings of the International Conference Recent Advances in Natural Language Processing*, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pp. 672–680. https://aclanthology.org/R15-1086.

Vidgen, Bertie, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts (2019), Challenges and frontiers in abusive content detection, Association for Computational Linguistics.

Vidgen, Bertie and Leon Derczynski (2020), Directions in abusive language training data, a systematic review: Garbage in, garbage out, *Plos one* **15** (12), pp. e0243300, Public Library of Science San Francisco, CA USA.

Waseem, Zeerak, Thomas Davidson, Dana Warmsley, and Ingmar Weber (2017), Understanding abuse: A typology of abusive language detection subtasks, *arXiv preprint arXiv:1705.09899.*

y Jorge Carrillo-de-Albornoz y Laura Plaza y Julio Gonzalo y Paolo Rosso y Miriam Comet y Trinidad Donoso, Francisco Rodríguez-Sánchez (2021), Overview of exist 2021: sexism identification in social networks, *Procesamiento del Lenguaje Natural* **67** (0), pp. 195–207. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

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

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar (2019a), Predicting the type and target of offensive posts in social media, *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, pp. 1415–1420. https://aclanthology.org/N19-1144.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar (2019b), Predicting the type and target of offensive posts in social media, *arXiv preprint arXiv:1902.09666.*

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar (2019c), SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 75–86. https://aclanthology.org/S19-2010.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar (2019d), Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), *arXiv preprint arXiv:1903.08983.*

Zhou, Yucheng, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang (2021), Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 5822–5834. https://aclanthology.org/2021.naacl-main.465.

Appendix A.

Preprocessing All experiments have been conducted with common pre-processing steps, namely:

- lowercasing of all words
- all users' mentions have been substituted with a placeholder (@USER);
- all URLs have been substituted with a with a placeholder ([URL]);
- all ordinal numbers have been replaced with a placeholder ([NUM]);
- emojis have been removed;
- hashtag symbol has been removed from hasthtags (e.g. #kadiricinadalet → kadiricinadalet);
- extra blank spaces have been replaced with a single space;
- extra blank new lines have been removed.

Appendix B.

		dyna-BERTje-Silver		
Training iterations	**Class**	**Prec.**	**Rec.**	**Macro F1**
0	NOT	0.736	0.775	0.643
	OFF	0.558	0.506	
1	NOT	0.731	0.808	0.643
	OFF	0.578	0.469	
2	NOT	0.741	0.817	0.659
	OFF	0.602	0.491	
3	NOT	0.749	0.809	**0.667**
	OFF	0.603	0.517	
4	NOT	0.753	0.789	0.666
	OFF	0.589	0.538	
5	NOT	0.748	0.794	0.661
	OFF	0.587	0.523	
6	NOT	0.740	0.787	0.650
	OFF	0.572	0.506	
7	NOT	0.757	0.769	0.665
	OFF	0.576	0.560	
8	NOT	0.758	0.743	0.658
	OFF	0.557	0.577	
9	NOT	0.751	0.776	0.660
	OFF	0.575	0.541	
10	NOT	0.743	0.774	0.650
	OFF	0.565	0.523	

Table 14: Results for the dyna-mBERT-PSP and mBERT-PSP+ over the full 10 iterations on OP-NL. Macro-average F1 is used to identify the best performing models. Precision and Recall per class are reported as well. Best results for each model are marked in bold.