MAD-TSC: A Multilingual Aligned News Dataset for Target-dependent Sentiment Classification

Evan Dufraisse[†*], **Adrian Popescu**[†], **Julien Tourille**[†], **Armelle Brun**[*], **Jerome Deshayes**[†]
[†] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[*] Université de Lorraine - CNRS - Loria, Vandoeuvre les Nancy Cedex, France

Abstract

Target-dependent sentiment classification (TSC) enables a fine-grained automatic analysis of sentiments expressed in texts. Sentiment expression varies depending on the domain, and it is necessary to create domain-specific datasets. While socially important, TSC in the news domain remains relatively understudied. We introduce MAD-TSC, the first multilingual aligned dataset designed for TSC in news. MAD-TSC differs substantially from existing resources. First, it includes aligned examples in eight languages to facilitate a comparison of performance for individual languages, and a direct comparison of human and machine translation. Second, the dataset is sampled from a diversified parallel news corpus, and is diversified in terms of news sources and geographic spread of entities. Finally, MAD-TSC is more challenging than existing datasets because its samples are more complex. We exemplify the use of MAD-TSC with comprehensive monolingual and multilingual experiments. The latter shows that machine translations can successfully replace manual ones, and that performance for all included languages can match that of English by automatically translating test examples.

1 Introduction

Text analysis needs to address both *objective* aspects, such as topic extraction, and *subjective* aspects, such as sentiment and opinion classification. In spite of recent progress brought by the introduction of large language models (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019), sentiment classification remains a challenging task. Expression of sentiments varies according to the data sources, languages and domain of the texts. These challenges are particularly important in target-dependent sentiment classification (TSC), which focuses on determining the sentiment expressed toward a given entity in a given context. The bulk of TSC-related research efforts are focused on major languages. This focus is an effect of the availability of generic and of task-specific resources in these languages (Brauwers and Frasincar, 2022; Nazir et al., 2020). A majority of datasets are monolingual and, when they are multilingual (Balahur and Turchi, 2014; Barriere and Balahur, 2020; Cortis and Davis, 2021; Pontiki et al., 2016; Severyn et al., 2016), the examples are not aligned across languages. Equally, a majority of existing datasets and methods are devised for texts such as tweets, reviews or comments (Nakov et al., 2016; Pontiki et al., 2016; Severyn et al., 2016) in which sentiment is most often expressed explicitly. Somewhat surprisingly, less attention is given to TSC in news, despite the usefulness of an automatic analysis of their content for the understanding of societally impactful processes such as disinformation or polarization (Hamborg and Donnay, 2021).

Our main contribution is the introduction of MAD-TSC, the first large multilingual aligned dataset for TSC in news articles. It includes 5,110 annotated entity mentions from 4,714 unique sentences. Each sentence has professionally-translated and aligned versions in eight languages (English, Spanish, German, Italian, French, Portuguese, Dutch, and Romanian). Sentences originate from 286 news sources published in over 30 countries. These characteristics differentiate the proposed dataset from existing resources, and in particular from NewsMTSC (Hamborg and Donnay, 2021), a monolingual dataset focused on American politics which is the closest to MAD-TSC. We first present the dataset creation process, and provide a qualitative analysis of its content. Then, we propose a thorough evaluation of state-of-the-art TSC methods on MAD-TSC in monolingual and multilingual settings, with particular focus on the usability of machine translation in TSC. The main findings are the following:

- The proposed dataset is more challenging since

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8286–8305

it includes more complex examples compared to existing resources, as detailed in Subsection 3.5.

- Performance for individual languages varies due to the fact that the quality of available pretrained language models is itself variable, with the best scores being obtained for English.
- Results with machine translation of training sets from English toward target languages and with manual translations are on par.
- The performance level for other languages matches that of English by translating test examples to English in order to take advantage of the pretrained language models available in this language. The same is true when both the training and test sets are automatically translated to English.

The last two findings are particularly interesting since they show that if a domain-specific TSC dataset is available, it can be effectively used for multiple languages. Overall, the introduction of MAD-TSC will facilitate progress in multilingual TSC. The dataset and the full evaluation protocol are available online[1] to encourage further research, and to ensure reproducibility.

2 Related work

Target-dependent sentiment classification (also named aspect-based sentiment analysis (Nazir et al., 2020) or classification (Brauwers and Frasincar, 2022)) is a complex task due to the numerous factors which influence the way sentiments are expressed in texts (Brauwers and Frasincar, 2022), such as the language, the domain or the personal biases of the author/reader.

TSC is often evaluated on short texts such as tweets (Nakov et al., 2016), reviews (Pontiki et al., 2016) or comments (Severyn et al., 2016). A common characteristic of these resources is that sentiment is often expressed in an explicit way. News are more challenging texts because sentiments are expressed implicitly or indirectly (Hamborg and Donnay, 2021), often include multiple targets in a single sentence (Brauwers and Frasincar, 2022), and both negative and positive arguments about a target entity are combined due to the fact that journalists are supposed to be objective (Balahur et al., 2010; Hamborg et al., 2019; Liu, 2010).

Multilinguality is important in order to be able to analyze texts in different languages. Multilingual datasets are proposed for tweets (Lampert and Lampert, 2021; Vilares et al., 2017), reviews (Jiménez Zafra et al., 2015; Pontiki et al., 2016) and institutional texts (Cortis and Davis, 2021). While interesting, these datasets differ from MAD-TSC because they do not focus on news. Equally important, they are only aligned at a domain level, but not at an example level. This limits their utility in terms of comparative evaluation in monolingual and multilingual settings. Multilingual approaches were also explored for news representation. For instance, bilingual word embeddings were used to compensate data scarcity in under-resourced languages (Akhtar et al., 2018) or to transfer models between source and target languages in zero-shot settings (Jebbara and Cimiano, 2019).

Classical TSC methods rely on engineered features based on lexicons and syntactic analysis (Biber and Finegan, 1989; Baccianella et al., 2010; Jiang et al., 2011; Kiritchenko et al., 2014; Vo and Zhang, 2015). Strong progress in TSC was made possible by the introduction of deep language models, such as BERT (Devlin et al., 2019; Zeng et al., 2019). Improvements are obtained when pretraining includes a larger proportion of news (Hamborg and Donnay, 2021) – this is the case for English RoBERTa (Liu et al., 2019) or XLNET (Yang et al., 2019) – or when including an intermediate tuning on domain-related data (Du et al., 2020). Naturally, further improvements are obtained by introducing TSC-specific architectures (Brauwers and Frasincar, 2022; Nazir et al., 2020; Zhou et al., 2019). We follow this trend and use pretrained language models in our experiments.

To our knowledge, there are only four datasets dedicated to TSC in news. The first one (Balahur et al., 2010) has 1,592 examples, and includes only quotes from political news. Quotes are interesting because they include a lot of sentiment expressions, but they are also easier since sentiment is often expressed explicitly (Hamborg and Donnay, 2021). The second one (Steinberger et al., 2017) has 1,274 examples. The size of these datasets makes them difficult to use with deep-learning-based TSC methods. The third dataset (Hamborg et al., 2019) includes 3,002 examples. However, as noted later by its authors (Hamborg and Donnay, 2021), the dataset is imbalanced and its sentiment expressions are predominantly explicit. The dataset which is closest to ours is NewsMTSC (Hamborg and Donnay, 2021). Their common characteris-

[1] https://github.com/EvanDufraisse/MAD_TSC

tics include: a focus on political news, an identical definition of the task, and a similar annotation process. Importantly, we follow Hamborg and Donnay (2021) and instruct annotators to think from the author's perspective in a holistic way. They are instructed to consider the "what" of the sentence (events, facts) but also the "how" (choice of words, author's attitude). This choice contrasts with previous works (Balahur et al., 2010; Steinberger et al., 2017), which distinguish author- and reader-levels in TSC, and is important in order to minimize the influence of personal biases. The main differences between MAD-TSC and NewsMTSC relate to multilingualism, complexity of examples, political topics and geography of examples, while maintaining the same order of magnitude in the number of samples, with 5,110 for MAD-TSC , and 11,361 for NewsMTSC. These differences are detailed in Subsection 3.5, and they make MAD-TSC appropriate for a thorough evaluation of TSC in multilingual settings.

3 Dataset Construction

We build on previous works devoted to the creation of sentiment classification datasets (Nakov et al., 2016; Pontiki et al., 2016), and particularly of TSC ones (Balahur et al., 2010; Hamborg and Donnay, 2021; Steinberger et al., 2017). We describe the main steps of the dataset creation process and analyze the resulting dataset.

3.1 Data Sources

Our objective is to create a dataset which: (1) is multilingual and aligned at a sentence level across languages to enable a comprehensive evaluation of TSC, (2) includes content from a large number of high-quality journalistic sources, which offer a diversified view of the included topics, (3) covers societally-impactful topics in different countries. Voxeurop[2] is a multilingual news website which aims to offer interesting and high-quality news to European audiences. The project inherits from Presseurop, which was active from 2009 to 2013, and whose objective was to make Europe-related news from over 200 sources available. Voxeurop and Presseurop articles are available in up to ten languages. The content is translated by professional translators, thus ensuring high-quality texts in all available languages. The content is published using a Creative Commons BY-NC-ND, an open license which facilitates its redistribution and reuse for non-commercial purposes, which will be also used to distribute MAD-TSC. We have collected 7,370 news articles which have translations in all eight languages, amounting to a total of 122,263 sentences in English and comparable numbers in other languages. A wide majority of the examples are related to politics, with the others pertaining to business, culture and society. Most of the entities mentioned in articles refer to prominent public figures from different European countries at the time of publication (2009–2013). Well-represented political sub-topics include: the economic crisis which started in 2008, European Union evolution process, election-related news, political crisis at a national level, and corruption scandals in different countries.

3.2 Sample Selection

Named entity detection was performed using the Flair model for English (Akbik et al., 2018), which led to an initial pool of 30,303 sentences with at least one mention of a person. We combine entity linking with Blink (Wu et al., 2020) and coreference resolution with neuralcoref[3] to obtain reliable entity counts in articles. Entities which are mentioned a single time in an article are not kept for annotation because they are not considered in focus. This filtering led to 19,223 candidate sentences. The alignment of sentences for the eight languages is inspired by lingtrain[4]. It is based on sentence embeddings from a multilingual-sentence BERT model (Yang et al., 2020), with English as source and the other languages as targets. Two matching criteria are used: (1) the need for reciprocal best matching (inter-match) between source and target sentences, and (2) a cosine similarity threshold of 0.5. Both criteria need to be met for all language pairs in order to select a sentence. Automatic alignment was manually checked for three languages (EN, FR, RO), with a sample of 1,000 examples. It was correct in 98.1% of cases. The remaining imperfections usually relate to additional context being provided by the translator in one of the languages. This does not affect the sentiment expressed about the target entity, and the obtained alignments of sentences are usable.

Following sentence alignment, it is also necessary to align entity mentions across languages, and

[2]https://voxeurop.eu

[3]https://github.com/huggingface/neuralcoref

[4]https://github.com/averkij/lingtrain-aligner

we used a rule-based approach for this task. NFKD unicode normalization is first applied to examples in all languages. Then we computed a normalized Levenshtein distance between the English mention of the entity and the words from the target sentence. A similarity threshold of 80% between the English and the target mention in any of the other languages was used. To add flexibility to the matching process, we also considered nearly contiguous sequences as valid matches. We have checked this matching and it is correct in over 99% of cases on a subset of 1,000 mentions.

The sentiment classes are not evenly distributed in news, and we followed an initial selection procedure which is inspired by the one introduced in Hamborg and Donnay (2021). It involves an undersampling of potentially neutral mentions as predicted by a simple binary classifier. This led to a pool of 11,000 examples which were selected and proposed to annotators.

3.3 Sample Annotation

Annotations were crowdsourced using a custom web interface. Aware of the challenges of news annotations (Balahur et al., 2010) (i.e. low inter-annotator agreement and low suitability), Hamborg and Donnay (2021) devised a process in which participants are asked to annotate following the news author's perspective. While some news articles express sentiments in a manner which is easy to recognize, others convey them in an intricate and/or implicit manner. Equally, the sentiment often depends on text parts distant from the target entity, which are not included in the text presented for annotations. The annotation guide made the annotators aware of the complexity of the task. They were presented with examples of sentences which include intricate and/or implicit sentiments expressions, as well as irony. Annotators also had the possibility to label examples as unknown whenever they could not determine the label of an example.

TSC annotations are usually collected using 3-, 5- or 7-points Likert scales (Balahur et al., 2010; Hamborg and Donnay, 2021; Nakov et al., 2016; Pontiki et al., 2016). Following an initial experiment which involved 50 sentences, we opted for a 5-points scale which offers a good balance between annotation simplicity and expressiveness. Possible labels were: negative, weakly negative, neutral, weakly positive and positive. The annotation was supported by a Web interface which is illustrated in Appendix E. Examples were provided in English, French and Romanian to facilitate the annotation.

Annotations were provided by a total of 21 volunteer participants, whose demographics are presented in Appendix D. They were recruited via a call for participation which was circulated via group and personal e-mails. Participants provided explicit consent to use their annotations and demographic data at the beginning of the experiment. The choice to work with volunteers is motivated by the fact that crowdsourcing performance is similar for paid and volunteered participation (Mao et al., 2013). Samples were presented randomly in order to avoid any ordering effect, and users were free to stop at any point. Each sample was labeled by three annotators in order to allow annotation consolidation. All users speak at least two of the three languages used in the annotation interface.

3.4 Label Aggregation

Since the annotation task is prone to disagreement (Hamborg and Donnay, 2021; Steinberger et al., 2017), a consolidation of annotations is necessary. We first removed all samples for which there was at least one "unknown" label. Following Hamborg and Donnay (2021), we reduced the five initial labels to three classes (negative, neutral and positive) by aggregating the two possible labels for the negative and positive sentiments. Finally, we kept only samples for which there was an unanimous voting or majority agreement with a third vote in a neighboring class (for instance two positive, one neutral). The inter-rater reliability, measured using Fleiss' kappa (Fleiss, 1971), reaches $\mathcal{K}_F = 0.58$ and $\mathcal{K}_F = 0.67$, before and after consolidation, respectively. The two values indicate that the task is challenging, but the final reliability score corresponds to good agreement (Hallgren, 2012). This consolidation strategy leads to a coherent labeling of MAD-TSC, and is used in experiments.

3.5 MAD-TSC Analysis

MAD-TSC includes a total of 5,110 labeled target entity mentions for all eight languages, with 1,839, 2,011, 1,260 of them labeled as negative, neutral and positive, respectively. There can be more than one target entity per example, and there are 4,714 unique examples in MAD-TSC.

Figure 1 shows the total number of unique linkable entities, grouped by source language, from the original corpus of articles used to design MAD-

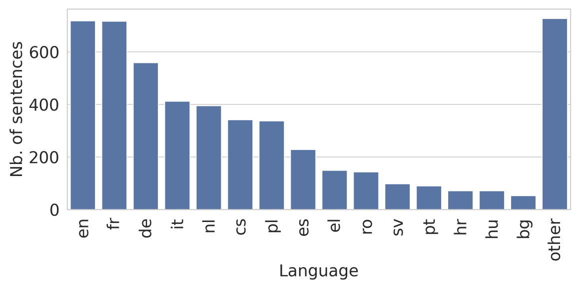


Figure 1: Distribution of linkable entities per original language of source article. "other" merges rare source languages. Linking is done with Blink (Wu et al., 2020).

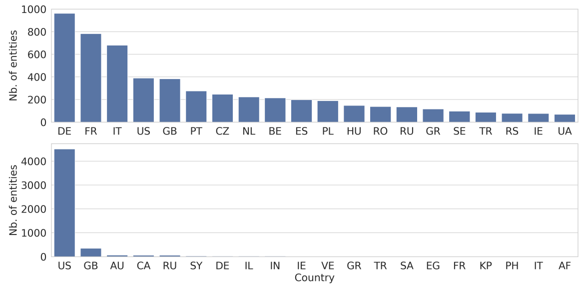


Figure 2: Distribution of linkable entities from MAD-TSC (top) and NewsMTSC (bottom) per country. The country associated to each entry is the most frequent one from the Wikipedia article of the entity.

TSC . These sources are generally correlated to journals from different countries, indicating that the proposed dataset comes from a variety of sources. The total number of unique linkable entities aggregated across all source languages is lower, with 1,007 entities. This is in contrast with NewsMTSC, which is sampled only from American newspapers.

We illustrate the content of the dataset by a number of quantitative and qualitative characteristics. We first present example-related statistics for MAD-TSC, and compare them to NewsMTSC (Hamborg and Donnay, 2021). These authors use sentence length as a proxy for the complexity of the dataset, and showed that texts in NewsMTSC are longer than those from previous datasets. We follow their approach and report the statistics regarding the number of characters for English examples included in MAD-TSC. The mean is 192.3 characters (72.1 stdev), while the corresponding value for NewsMTSC is 152.2 (109.1 stdev). The number of words per example is a related measure, but more informative from a semantic perspective. MAD-TSC examples include a mean of 31.1 words (11.6 stdev), while the corresponding values are 25.2 (15.5 stdev) for NewsMTSC. MAD-TSC can thus be considered as more complex than NewsMTSC.

A second analysis focuses on entities which can be linked to English Wikipedia articles using Blink (Wu et al., 2020). MAD-TSC contains 1,007 distinct linkable entities, with a mean of 6.3 mentions per entity (28.9 stdev). NewsMTSC includes 975 linkable entities, 5.5 mentions per entity (43.9 stdev). Donald Trump, Hillary Clinton, and Barack Obama, the top-3 entities from NewsMTSC cover 20.1%, 11.7% and 7.7% of mentions, respectively. In MAD-TSC, Angela Merkel, Silvio Berlusconi, and Nicolas Sarkozy cover 10.2%, 4.9%, and 3.5% of mentions, respectively. We conclude that MAD-TSC has a distribution of entity mentions which is less skewed.

Third, we examine the geographic distribution of linkable entities in both datasets. The obtained distributions are presented in Figure 2, and they confirm that MAD-TSC is much more diversified than NewsMTSC from a geographical point of view.

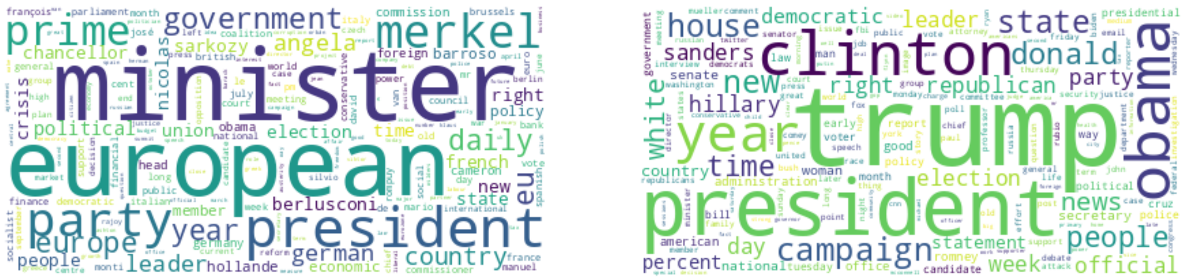
Finally, we present clouds of frequent words in the two datasets in Figure 3. This qualitative representation of the two datasets confirms that the main topics are different, with focus on European and on American topics, respectively.

4 Experiments

4.1 Compared TSC Methods

For English, we compare the performance of four existing TSC methods using English and multilingual pretrained language models. Whenever available, RoBERTa (Liu et al., 2019) backbone is preferred due to its better performance for TSC in news (Hamborg and Donnay, 2021). Otherwise, a BERT (Devlin et al., 2019) backbone is selected. Details about pretrained models used in experiments are provided in Appendix A. A short description of the TSC methods is provided below:

- SPC (Song et al., 2019) is based on the classical Sentence Pair Classification task of Bert models. The input is designed as “[CLS] <sentence> [SEP] <entity> [SEP]”.
- TD (Gao et al., 2019) only considers the last hidden states of the entities’ tokens and merges their representation through a max-pooling layer.
- PM (Seoh et al., 2021) employs a prompt model for TSC. Our implementation of this method uses the simple prompt "<entity> is [mask]" with ["good", "ok", "bad"] proposed in the original paper as verbalizer. This prompt is translated in



(a) MAD-TSC dataset

(b) NewsMTSC dataset

Figure 3: Frequent word clouds for MAD-TSC and NewsMTSC corpora (minimum frequency 20 occurrences).

all MAD-TSC languages.

- BASE - version of SPC without access to the entity mention in its input. It can be deployed for any general sentiment classification strategy since annotations of the entity are not used.

We implement the efficient fine-tuning process introduced by Mosbach et al. (2021) to optimize our TSC models. This type of optimization was successfully used in a TSC context (Seoh et al., 2021). We fix the learning rate to 2e-5 and train with early-stopping conditions up to 40 epochs using AdamW optimizer, and batches of size 32. Initial experiments confirmed that this approach is better than the more classical hyperparameter search used in BERT (Devlin et al., 2019). Details about the optimization process are provided in Appendix B. All results are reported by averaging scores of five runs launched with different seeds. Pytorch (Paszke et al., 2019) is used for all implementations.

4.2 Dataset Splits and Metrics

We run experiments with MAD-TSC in monolingual and multilingual settings, and also use NewsMTSC for experiments in English. The training/validation/test subsets are sampled randomly and include 3,810/300/1,000 labeled mentions, respectively. Results for NewsMTSC are reported with the official splits from Hamborg and Donnay (2021). TSC evaluation can be performed with different metrics (Hamborg and Donnay, 2021; Nakov et al., 2016; Pontiki et al., 2016), we use macro F1 ($F1_m$) on all classes as primary metric. Performance with other metrics, such as F1 only on positive and negative classes ($F1_{pn}$), accuracy (acc), and average recall (rec) follow similar trends. A selection of such results, along with standard deviations are reported in Appendix C.

Train	NM		MAD		MIX	
Test	NM	MAD	NM	MAD	NM	MAD
SPC$_{EN}$	83.2	67.1	76.7	72.3	**83.4**	72.7
SPC$_{ML}$	81.0	61.5	70.5	67.8	80.3	67.9
TD$_{EN}$	**83.6**	69.7	75.2	**73.2**	83.0	**74.1**
TD$_{ML}$	81.6	63.3	71.9	68.5	80.4	68.4
PM$_{EN}$	**83.6**	**67.3**	**77.3**	72.8	82.6	72.2
PM$_{ML}$	81.7	60.9	69.3	66.6	82.0	67.0
BASE$_{EN}$	76.7	64.0	68.8	69.9	76.7	71.0
BASE$_{ML}$	74.1	59.2	66.5	64.8	74.1	66.7

Table 1: $F1_m$ results for individual train and test sets of the English subset of MAD-TSC (MAD) and NewsMTSC (NM), and for their combination (MIX). TSC methods are applied on top of language models pretrained specifically for English (EN) or with a multilingual corpus (ML).

4.3 Experiments with MAD-TSC and NewsMTSC

We compare the TSC methods from Subsection 4.1 on both the English subset of MAD-TSC and on NewsMTSC. Models are trained and evaluated on each dataset and on their combination. Results with different train and test set combinations are reported in Table 1. Regardless of the specific combination, $F1_m$ scores obtained with SPC, TD and PM are similar. The fact that BASE has lower performance (approximately 6 and 3 points for NewsMTSC and MAD-TSC, respectively) confirms that the need to provide the target entity in TSC. The performance of SPC and TD obtained when training and testing on NewsMTSC is 3 to 4 points better than the one originally reported in (Hamborg and Donnay, 2021). This is probably due to a better parametrization of these two methods here. SPC and TD scores are even on a par with those of the more complex GRU method (Hamborg and Donnay, 2021), which needs an external knowledge source that is not available for languages other than English. In addition, results confirm that MAD-

Pretrain	EN	ES	DE	IT	FR	PT	NL	RO
TG	**72.3**	63.9	**66.1**	**65.8**	**70.8**	**68.2**	62.1	66.9
ML	67.8	**67.2**	64.8	65.1	67.2	66.2	**66.4**	**68.5**

Table 2: $F1_m$ results for the eight languages included in MAD-TSC. SPC is applied on top of models pretrained specifically for each target language (TG) or with a multilingual corpus (ML) using SPC.

TSC is more challenging than NewsMTSC. The transfer of the models trained on one dataset toward the other test set gives suboptimal results. The combination of the two train sets has a marginal positive effect on each test set. Finally, the language model pretrained specifically for English is clearly better than its multilingual counterpart, probably due to the curse of multilinguality which affects multilingual models (Conneau et al., 2020).

4.4 Experiments with Individual Languages

Results for the eight MAD-TSC languages are presented in Table 2. They are reported using SPC, a commonly used TSC method (Cao et al., 2022; Hamborg and Donnay, 2021; Seoh et al., 2021; Song et al., 2019), and its performance is close to that of PM and TD in Table 1. The best $F1_m$ scores are obtained for English and French, and the lowest scores are reported for Dutch and Spanish. When using monolingual pretraining (TG), the difference between the best and worst scores is over 10 points. In contrast, the results obtained with multilingual pretraining (ML) are much more similar across all languages. Performance variability is explained by the quality of pretrained models, and in particular by the size of the datasets and that of the subsets relevant for politics. Interestingly, the multilingual pretraining is much better than the language-specific one for Dutch and Spanish, and is also better for Romanian. Inversely, monolingual models are clearly better for English and French, that are the two languages which have the best monolingual pretraining. The results from Table 2 indicate that strong monolingual pretraining is preferable in TSC, but it can be successfully replaced by multilingual pretraining when the dataset for a particular language is insufficient.

4.5 Experiments with Machine Translation

Machine translation (MT) has strongly progressed in recent years, notably due to the introduction of neural architectures (Stahlberg, 2020). A successful deployment of MT for sentiment classification

Train	Test	ES	DE	IT	FR	PT	NL	RO
EN	EN_{M2M}	73.3	70.8	71.4	70.6	71.9	71.1	73.0
EN	EN_{DL}	**73.9**	**73.2**	**72.5**	**73.5**	**73.1**	**72.1**	**73.8**
TG	TG	*63.9*	*66.1*	*65.8*	*70.8*	*68.2*	*62.1*	*66.9*
TG_{M2M}	TG	**64.7**	65.0	**66.0**	70.6	66.9	**64.2**	65.7
TG_{DL}	TG	63.7	**65.2**	65.8	**71.3**	**68.3**	62.0	**67.5**

Table 3: $F1_m$ results for machine translation languages included in MAD-TSC, compared to the results obtained when without machine translation for English-only (72.3 in Table 1) and monolingual models (fourth row copied from Table 2). Notations: EN - English, TG - target language. The original train/test sets were used if no subscripts are present. DL (DeepL) and $M2M$ (Fan et al., 2021) subscripts give the machine translation model used. All results are reported with language-specific pretrained models. TSC models are trained with SPC.

would greatly facilitate the task in the multilingual setting because it would reduce, or even remove, the need for specific annotations in each language. Building on previous works which apply MT to TSC (Balahur and Turchi, 2014; Mohammad et al., 2016), we report results with English as pivot language. Test and/or train subsets of the other languages are translated to English. The translation is performed with two methods: (1) M2M100 (Fan et al., 2021), a recent massively multilingual translation model, by using the largest available model (12B parameters); (2) the API of DeepL[5], a well-known commercial machine translation service.

The $F1_m$ scores obtained with different MT configurations are reported in Table 3. The results are very interesting, particularly when translating the test set to English with DeepL (row with EN train and EN_{DL} test). $F1_m$ scores are globally better than 72.3, the performance of SPC obtained with manual translations for English in Table 1. The maximum gain is 1.6 points for Spanish, and Dutch is the only language for which DeepL translations are slightly worse (-0.2 points). $F1_m$ is also interesting with M2M100, albeit lower than that of EN_{DL}. These results surprised us initially, but we reach the same conclusions after running the experiments a second time in an independent manner. When translating the test set, all languages benefit from the strong pretraining available for English, and strong performance can be obtained for them if a good translation model is available from the target language toward English. This condition is met for all languages included in MAD-TSC . Qualitatively, the findings reported here could be

[5] https://www.deepl.com/en/docs-api

explained by the fact that, while the polarity of sentiments is preserved by both machine and human translators, professional translators are more creative and sometimes add context in sentences for their international public. While useful for human readers, these changes can have a slight negative influence on sentiment classification.

Results are also interesting when the English training set is translated toward the target languages using DeepL and M2M100 (rows with TG_{DL} / TG_{M2M} train and TG test). The associated $F1_m$ scores are on par with those obtained with the manual translations. However, the translation of training sets is less effective than that of test sets. This happens because TSC training is done in languages other than English, and is based on weaker pretrained language models.

We also translated both the training and test sets from Dutch and Romanian to English, two of the languages which have low performance in Table 2, using DeepL. The $F1_m$ scores for the two languages are 72.3 and 73.6, respectively. We conclude that interesting performance can be obtained by automatically translating TSC datasets from other languages to English.

4.6 Analysis of Example Complexity

We complement the quantitative results by a qualitative analysis of factors which influence TSC performance. The analysis is done for English, and findings are similar for the other languages.

We first test the hypothesis that the complexity of examples is correlated to their length (Hamborg and Donnay, 2021). We split sentences in five subsets depending on the number of words per example and report $F1_m$ per subset: 75.0 for up to 20 words per sentence; 72.9 for 21 to 30 words; 70.9 for 31 to 40 words; 70.0 for 41 to 50 words; 70.0 for 50 words and more. These scores confirm that TSC difficulty increases with example length, but differences become smaller above 30 words.

The number of entities detected in each example is an interesting proxy for complexity since the expressed sentiments can vary for multiple entities. We compute $F1_m$ separately for examples which include 1, 2, 3 or more detected entities. The scores obtained for the three subsets are 73.6, 70.1, and 67.4, respectively. They confirm that the presence of more entities makes TSC more difficult.

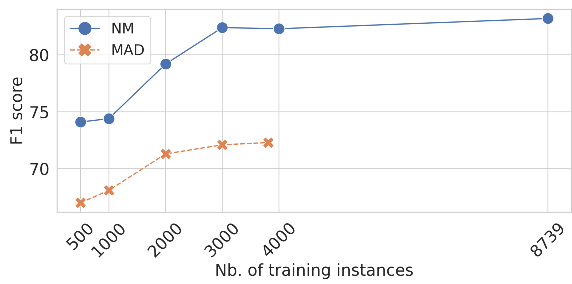


Figure 4: $F1_m$ scores with training sets of variable size for NewsMTSC (NM) and MAD-TSC (MAD). Results are reported with monolingual pretraining and SPC. The rightmost points represent scores for the full dataset.

4.7 Ablation of Training Set

The annotation of TSC datasets is a cumbersome task, and it is important to minimize this effort, while preserving the final performance. In Figure 4, we present the results obtained by sampling 1,000, 2,000, 3,000 and 4,000 training examples from NewsMTSC and MAD-TSC, and with the full datasets. $F1_m$ scores increase up to 3,000 samples, but the added benefit of more samples is reduced beyond this dataset size. The trend is similar for MAD-TSC , and the results reported in Figure 4 indicate that its size is sufficient to tackle the TSC task effectively.

5 Conclusion

We introduce MAD-TSC, a dataset for multilingual target-dependent sentiment classification. Compared to existing resources, the proposed dataset is aligned across languages, includes examples about geographically diversified entities. Examples are longer and more complex because sentiment is often expressed in an implicit way. Given its aligned character, MAD-TSC dataset enables a comparison of sentiment classification between languages. Performance varies significantly, and this variation is to a large extent explained by the quality of pretrained models available for each language.

Importantly, the MT experiments show that human translations can be replaced by automatic ones. The automatic translation of test sets from target languages to English is particularly interesting since it brings target-dependent sentiment classification in different languages to the same quality level as that of English. This allows TSC to be scaled for the languages included in this study without the need to develop language-specific training sets. The only condition is to have a labeled domain-related dataset in one language, English or other.

Future work will focus on limitations of the dataset: (1) the handling of example complexity, (2) the coverage of the dataset in terms of domains and entity types, (3) the number of included languages, and (4) the quality of pretrained language models. These limitations are discussed in more details in the dedicated section below.

Limitations

The analysis from Subsection 4.6 shows that the number of entities per example has an important influence on results. For now, the handling of examples with more than one entity includes the detection of the mention, but does not consider other criteria. It would be useful to adapt the TSC methods in order to determine whether the sentiment about all entities is expressed by the same part of the example or not. If sentiment is expressed in different parts of the example, a splitting of the example into parts which express the sentiment about each entity would prove useful.

Despite a more diversified coverage of the political domain compared to NewsMTSC, MAD-TSC remains focused on politics. It would be interesting to include other major news domains, such as environment, business, culture, technology, sports, etc. Equally important, all targets from MAD-TSC are person names. It would be useful to also include sentiment expressed about other types of named entities (organizations, locations, events, etc.), as well as other polarization-prone concepts in each domain. Such extensions of the dataset would provide a more complete view of TSC performance. Ultimately, they would make the analysis of sentiments expressed in news articles more comprehensive and reliable.

While MAD-TSC is the first multilingual aligned dataset designed for TSC in news, it would benefit from the inclusion of more languages, including non-European ones. This limitation is due to the unavailability of massively multilingual and aligned news datasets which could be used to include more languages. A potential solution to overcome these limitations would be to enrich the dataset with manual translations in other languages. However, this solution is costly and is left for future work.

Finally, the comparison of results between languages is biased because the effectiveness of available pretrained language models is variable. This is a limitation which is shared by most NLP works which focus on multilingual datasets and reuse pretrained models, themselves trained on whatever datasets available for each language.

Ethics Statement

The work presented here is part of a project which was reviewed and approved by our institution's ethical committee. This committee provided useful guidance regarding this specific work concerning the selection of news sources, and the annotation process. The recommendations were integrated in the dataset creation process.

One potential issue is that the dataset includes negative sentiments expressed about public figures. This is also applicable to any TSC dataset that focus on politics. Sentiment expressions were collected from publicly available news sources which have a right to freedom of expression in the European Union (EHCR Article 10). News articles were collected from diversified newspapers, which lean toward different parts of the political spectrum, and this reduces the risk of mischaracterizing any of the mentioned entities. Sentiment classification datasets are needed in order to understand how sentiment is expressed in the media, and thus contribute to the characterization of societal debates.

Acknowledgements

This work was supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media. This work has been funded by the BOOM ANR Project - ANR-20-CE23-0024. It was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.

Md Shad Akhtar, Palaash Sawant, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 572–582. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Valentin Barriere and Alexandra Balahur. 2020. Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 266–271. International Committee on Computational Linguistics.

Douglas Biber and Edward Finegan. 1989. Styles of stance in english: Lexical and grammatical marking of evidentiality and affect. *Text-interdisciplinary journal for the study of discourse*, 9(1):93–124.

Gianni Brauwers and Flavius Frasincar. 2022. A survey on aspect-based sentiment classification. *ACM Comput. Surv.*, 55(4).

Jiahao Cao, Rui Liu, Huailiang Peng, Lei Jiang, and Xu Bai. 2022. Aspect is not you need: No-aspect differential sentiment framework for aspect-based sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1599–1609. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Keith Cortis and Brian Davis. 2021. A dataset of multidimensional and multilingual social opinions for malta's annual government budget. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):971–981.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with BERT. *IEEE Access*, 7:154290–154299.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2486–2495. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160. Association for Computational Linguistics.

Salud M. Jiménez Zafra, Giacomo Berardi, Andrea Esuli, Diego Marcheggiani, María Teresa Martín-Valdivia, and Alejandro Moreo Fernández. 2015. A multi-lingual annotated dataset for aspect-oriented opinion mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2533–2538. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 437–442. Association for Computational Linguistics.

Jasmin Lampert and Christoph H Lampert. 2021. Overcoming rare-language discrimination in multi-lingual sentiment analysis. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5185–5192. IEEE.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. 2013. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *First AAAI conference on human computation and crowdsourcing*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18. Association for Computational Linguistics.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: a comprehensive survey. *IEEE Transactions on Affective Computing*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6311–6322. Association for Computational Linguistics.

Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2016. Multilingual opinion mining on youtube. *Information Processing & Management*, 52(1):46–60.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ralf Steinberger, Stefanie Hegele, Hristo Tanev, and Leonida Della Rocca. 2017. Large-scale news entity sentiment analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 707–715. INCOMA Ltd.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2017. Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3):595–607.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-fourth international joint conference on artificial intelligence*.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.

Jie Zhou, Jimmy Xiangji Huang, Qin Chen, Qinmin Vivian Hu, Tingting Wang, and Liang He. 2019. Deep learning for aspect-level sentiment classification: survey, vision, and challenges. *IEEE access*, 7:78454–78483.

A Language models

All models used in this work are publicly available in the Hugging Face models repository. The main characteristics of the pretrained models used in our experiments are presented in Table 4. Considering the improvements brought by RoBERTa (Liu et al., 2019) pretraining over subsequent fine-tuning tasks compared to BERT (Devlin et al., 2019), models that replicate RoBERTa in another language are favored when possible. All models are cased versions. Those models are between 100M and 150M of parameters.

Lang	architecture	link	reference
EN	roberta-base	RoBERTa-base	(Liu et al., 2019)
ES	roberta-base	Bertin v2	(De la Rosa et al., 2022)
DE	bert-base	bert-base german	-
IT	bert-base	bert-base italian	-
FR	roberta-base	Camembert	(Martin et al., 2020)
PT	bert-base	BERTimbau	(Souza et al., 2020)
NL	roberta-base	RobBERT	(Delobelle et al., 2020)
RO	bert-base	Romanian Bert	(Dumitrescu et al., 2020)
$MULTI$	roberta-base	XLM-RoBERTa	(Conneau et al., 2020)

Table 4: List of pretrained models, with associated language, base architecture, URL and reference when available.

The M2M model was downloaded from this repository[6].

B Training details

Hyperparam	Fine-tuning
Learning Rate	2e-5
Batch Size	32
Weight Decay	0.01
Warmup	0.06
Max Epochs	40
Adam ϵ	1e-6
β_1	0.9
β_2	0.98
Optimizer	AdamW
Seeds	[42, 302, 668, 745, 343]

Table 5: Values of hyperparameters used for training TSC models.

Choice of training hyperparameters have been made resulting from the reading of (Mosbach et al., 2021). However, we persist in the use of a validation set but add a loose early-stopping condition of 10 epochs without loss improvement. Results reported are for the model with the best validation loss found during training. Code will be made available upon publication for further implementation details.

The fine-tuning of a pretrained model for TSC takes approximately 3 hours on a Nvidia A100 card. We fine-tuned 52 models with 5 different seeds for each configuration. This leads to a budget of 780

[6]https://huggingface.co/facebook/m2m100-12B-last-ckpt

GPU-hours for training. The inference times for the various models used in this paper (Flair NER, Blink, translation models) amount to another 24 GPU-hours. The total budget spent for training and inference is 804 GPU-hours.

C Supplementary results

Tables 6, 7, 8 provide the standard deviation values for the experiments reported from Tables 6, 7, 8 of the main text. Notations from the main tables are preserved.

D Annotator Demographics

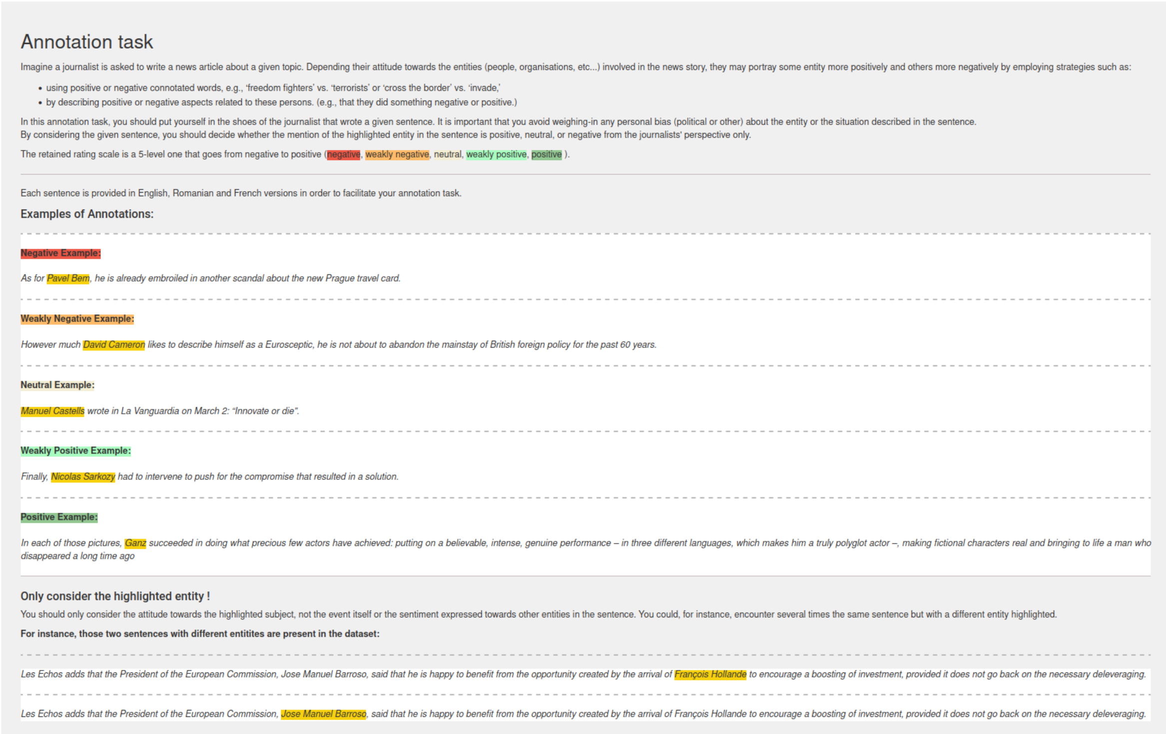
The main demographic characteristics of annotators are: (1) gender - 5 female/16 male; (2) age groups - 14 between 25 and 34, 5 between 35 and 44, 2 over 44, (3) countries of origin: France (16), Romania (4), Ivory Coast (1).

E Annotation interface

This section presents the top and bottom parts of the instructions and registration page (Figures 5 and 6, respectively), and an example of annotation page (Figure 7).

Model	Train	Test	$F1_m$	std_{F1_m}	$F1_{pn}$	$std_{F1_{pn}}$	acc	rec
SPC$_{EN}$	NM	NM	83.2	0.4	83.0	0.8	84.1	83.2
SPC$_{EN}$	NM	MAD	67.1	2.4	70.9	1.3	68.1	69.5
SPC$_{ML}$	NM	NM	81.0	0.3	80.3	0.5	81.6	81.1
SPC$_{ML}$	NM	MAD	61.5	1.3	64.7	1.3	61.8	63.2
TD$_{EN}$	NM	NM	83.7	0.7	83.4	0.8	83.8	83.0
TD$_{EN}$	NM	MAD	69.7	1.0	72.8	0.8	68.8	70.3
TD$_{ML}$	NM	NM	81.6	1.0	80.8	1.2	81.5	80.9
TD$_{ML}$	NM	MAD	63.3	1.5	65.5	1.4	62.9	64.3
PM$_{EN}$	NM	NM	83.6	0.6	83.1	1.0	84.1	83.5
PM$_{EN}$	NM	MAD	67.3	1.3	70.8	1.0	67.8	69.7
PM$_{ML}$	NM	NM	81.7	1.0	81.1	1.3	81.2	80.3
PM$_{ML}$	NM	MAD	60.9	1.9	63.7	1.3	61.7	62.5
BASE$_{EN}$	NM	NM	76.8	0.7	76.6	0.7	76.2	75.3
BASE$_{EN}$	NM	MAD	64.0	0.8	68.1	1.2	65.8	66.7
BASE$_{ML}$	NM	NM	74.1	0.7	74.0	1.2	74.6	74.3
BASE$_{ML}$	NM	MAD	59.2	1.5	62.8	0.8	60.5	62.3
SPC$_{EN}$	MAD	NM	76.7	1.0	76.6	1.7	77.3	75.7
SPC$_{EN}$	MAD	MAD	72.3	0.9	72.5	1.4	73.6	72.5
SPC$_{ML}$	MAD	NM	70.5	1.2	69.4	1.9	72.8	71.0
SPC$_{ML}$	MAD	MAD	67.8	1.6	66.9	1.7	68.6	68.0
TD$_{EN}$	MAD	NM	75.2	2.3	74.6	2.7	77.7	76.1
TD$_{EN}$	MAD	MAD	73.2	0.9	73.5	1.7	74.2	73.5
TD$_{ML}$	MAD	NM	71.9	1.2	70.0	2.8	75.0	72.6
TD$_{ML}$	MAD	MAD	68.5	1.1	66.6	2.0	69.8	68.5
PM$_{EN}$	MAD	NM	77.3	0.8	77.1	1.5	76.9	75.0
PM$_{EN}$	MAD	MAD	72.8	0.9	72.8	1.6	73.7	71.9
PM$_{ML}$	MAD	NM	69.3	4.8	67.3	6.5	75.0	73.5
PM$_{ML}$	MAD	MAD	66.6	3.1	64.7	4.6	69.3	68.8
BASE$_{EN}$	MAD	NM	68.8	0.6	68.6	1.4	69.4	68.2
BASE$_{EN}$	MAD	MAD	69.9	0.9	70.3	1.3	71.7	70.9
BASE$_{ML}$	MAD	NM	66.5	1.2	65.1	1.9	67.7	66.1
BASE$_{ML}$	MAD	MAD	64.8	1.1	62.9	2.2	66.9	65.9

Table 6: Scores and standard deviations of $F1_m$, $F1_{pn}$, an accuracy (acc) and macro averaged recall (rec) scores for different configurations train/test configurations of MAD-TSC (MAD) and NewsMTSC (NM). These results complement those presented in Table 1 of the main text.



Annotation task

Imagine a journalist is asked to write a news article about a given topic. Depending their attitude towards the entities (people, organisations, etc...) involved in the news story, they may portray some entity more positively and others more negatively by employing strategies such as:

- using positive or negative connotated words, e.g., 'freedom fighters' vs. 'terrorists' or 'cross the border' vs. 'invade,'
- by describing positive or negative aspects related to these persons. (e.g., that they did something negative or positive.)

In this annotation task, you should put yourself in the shoes of the journalist that wrote a given sentence. It is important that you avoid weighing-in any personal bias (political or other) about the entity or the situation described in the sentence.
By considering the given sentence, you should decide whether the mention of the highlighted entity in the sentence is positive, neutral, or negative from the journalists' perspective only.

The retained rating scale is a 5-level one that goes from negative to positive (negative, weakly negative, neutral, weakly positive, positive).

Each sentence is provided in English, Romanian and French versions in order to facilitate your annotation task.

Examples of Annotations:

Negative Example:

As for Pavel Bem, he is already embroiled in another scandal about the new Prague travel card.

Weakly Negative Example:

However much David Cameron likes to describe himself as a Eurosceptic, he is not about to abandon the mainstay of British foreign policy for the past 60 years.

Neutral Example:

Manuel Castells wrote in La Vanguardia on March 2: "Innovate or die".

Weakly Positive Example:

Finally, Nicolas Sarkozy had to intervene to push for the compromise that resulted in a solution.

Positive Example:

In each of those pictures, Ganz succeeded in doing what precious few actors have achieved: putting on a believable, intense, genuine performance – in three different languages, which makes him a truly polyglot actor –, making fictional characters real and bringing to life a man who disappeared a long time ago

Only consider the highlighted entity !

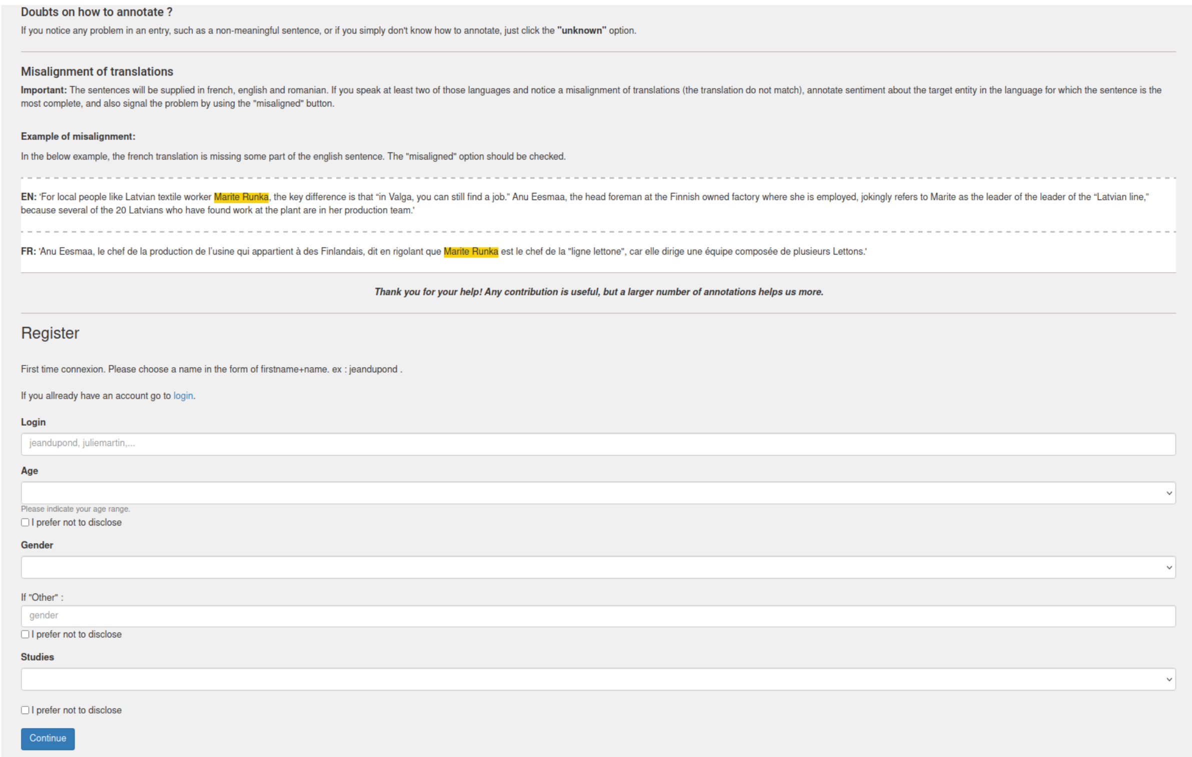
You should only consider the attitude towards the highlighted subject, not the event itself or the sentiment expressed towards other entities in the sentence. You could, for instance, encounter several times the same sentence but with a different entity highlighted.

For instance, those two sentences with different entitites are present in the dataset:

Les Echos adds that the President of the European Commission, Jose Manuel Barroso, said that he is happy to benefit from the opportunity created by the arrival of François Hollande to encourage a boosting of investment, provided it does not go back on the necessary deleveraging.

Les Echos adds that the President of the European Commission, Jose Manuel Barroso, said that he is happy to benefit from the opportunity created by the arrival of François Hollande to encourage a boosting of investment, provided it does not go back on the necessary deleveraging.

Figure 5: Top part of the instructions and registration page.



Doubts on how to annotate ?

If you notice any problem in an entry, such as a non-meaningful sentence, or if you simply don't know how to annotate, just click the **"unknown"** option.

Misalignment of translations

Important: The sentences will be supplied in french, english and romanian. If you speak at least two of those languages and notice a misalignment of translations (the translation do not match), annotate sentiment about the target entity in the language for which the sentence is the most complete, and also signal the problem by using the "misaligned" button.

Example of misalignment:

In the below example, the french translation is missing some part of the english sentence. The "misaligned" option should be checked.

EN: 'For local people like Latvian textile worker Marite Runka, the key difference is that "in Valga, you can still find a job." Anu Eesmaa, the head foreman at the Finnish owned factory where she is employed, jokingly refers to Marite as the leader of the leader of the "Latvian line," because several of the 20 Latvians who have found work at the plant are in her production team.'

FR: 'Anu Eesmaa, le chef de la production de l'usine qui appartient à des Finlandais, dit en rigolant que Marite Runka est le chef de la "ligne lettone", car elle dirige une équipe composée de plusieurs Lettons.'

Thank you for your help! Any contribution is useful, but a larger number of annotations helps us more.

Register

First time connexion. Please choose a name in the form of firstname+name. ex : jeandupond .

If you allready have an account go to login.

Login

jeandupond, juliemartin,...

Age

Please indicate your age range.

☐ I prefer not to disclose

Gender

If "Other" :

gender

☐ I prefer not to disclose

Studies

☐ I prefer not to disclose

Continue

Figure 6: Bottom part of the instructions and registration page.

Figure 7: Page for an example annotation.

Pretrain	$F1_m$	std_{F1_m}	$F1_{pn}$	$std_{F1_{pn}}$	acc	$recall$
EN_{TG}	72.3	0.9	72.5	1.4	73.6	72.5
EN_{ML}	67.8	1.6	66.9	1.7	68.6	68.0
ES_{TG}	63.9	2.0	61.9	3.5	66.1	63.9
ES_{ML}	67.2	1.1	64.6	1.9	69.0	68.1
DE_{TG}	66.1	1.0	64.6	1.2	67.3	65.8
DE_{ML}	64.8	1.3	62.9	2.1	66.1	65.1
IT_{TG}	65.8	1.0	64.8	1.6	67.0	65.6
IT_{ML}	65.1	1.2	63.2	1.9	66.2	65.3
FR_{TG}	70.8	0.8	69.6	1.6	72.5	71.5
FR_{ML}	67.2	1.2	65.6	2.4	69.2	68.2
PT_{TG}	68.2	0.3	66.7	1.3	69.1	67.6
PT_{ML}	66.2	1.3	63.9	1.8	67.8	66.9
NL_{TG}	62.1	3.8	60.1	5.4	64.9	64.1
NL_{ML}	66.4	1.6	64.9	2.8	66.8	65.2
RO_{TG}	66.9	1.1	66.4	1.2	68.0	66.2
RO_{ML}	68.5	1.1	68.0	1.8	70.8	69.7

Table 7: Scores and standard deviations of $F1_m$, $F1_{pn}$, an accuracy (acc) and macro averaged recall (rec) scores for monolingual experiments.These results complement those presented in Table 2 of the main paper.

Lang	Train	Test	$F1_m$	std_{F1_m}	$F1_{pn}$	$std_{F1_{pn}}$	acc	rec
ES	EN	EN_{M2M}	73.3	1.4	72.9	1.6	74.7	73.0
ES	EN	EN_{DL}	73.9	1.7	73.6	2.0	75.5	73.7
DE	EN	EN_{M2M}	70.8	1.5	70.1	1.6	72.2	70.6
DE	EN	EN_{DL}	73.2	1.0	72.6	1.4	74.7	72.9
IT	EN	EN_{M2M}	71.5	1.0	70.8	1.4	73.1	71.2
IT	EN	EN_{DL}	72.5	0.9	72.1	1.1	74.2	72.4
FR	EN	EN_{M2M}	70.6	1.3	70.0	1.6	72.6	70.4
FR	EN	EN_{DL}	73.5	1.3	73.3	1.5	75.2	73.3
PT	EN	EN_{M2M}	71.9	1.3	71.4	1.8	73.7	71.6
PT	EN	EN_{DL}	73.1	1.4	72.8	1.7	74.7	72.8
NL	EN	EN_{M2M}	71.1	1.3	70.7	1.6	72.7	70.7
NL	EN	EN_{DL}	72.1	0.9	71.8	1.2	74.0	71.8
RO	EN	EN_{M2M}	73.0	1.3	72.9	1.5	74.6	72.7
RO	EN	EN_{DL}	73.8	1.1	73.5	1.2	75.5	73.4
ES	TG	TG	63.9	2.0	61.9	3.5	66.1	63.9
ES	TG_{M2M}	TG	64.7	1.8	63.3	2.8	66.5	64.5
ES	TG_{DL}	TG	63.8	1.0	64.0	1.6	66.2	64.7
DE	TG	TG	66.1	1.0	64.6	1.2	67.3	65.8
DE	TG_{M2M}	TG	65.0	1.3	63.2	1.7	67.1	66.4
DE	TG_{DL}	TG	65.2	1.3	63.3	2.6	67.6	66.3
IT	TG	TG	65.8	1.0	64.8	1.6	67.0	65.6
IT	TG_{M2M}	TG	66.0	0.5	64.7	1.1	65.9	65.3
IT	TG_{DL}	TG	65.8	1.0	64.8	1.6	67.0	65.6
FR	TG	TG	70.8	0.8	69.6	1.6	72.5	71.5
FR	TG_{M2M}	TG	71.0	0.7	70.1	0.9	72.6	70.8
FR	TG_{DL}	TG	71.3	1.2	70.8	1.3	73.1	72.3
PT	TG	TG	68.2	0.3	66.7	1.3	69.1	67.6
PT	TG_{M2M}	TG	66.9	1.2	66.3	1.6	69.2	67.8
PT	TG_{DL}	TG	68.4	1.0	67.0	1.8	68.0	66.4
NL	TG	TG	62.1	3.8	60.1	5.4	64.9	64.1
NL	TG_{M2M}	TG	64.2	0.5	63.0	1.4	66.4	65.3
NL	TG_{DL}	TG	62.0	1.9	59.4	3.2	66.2	64.8
RO	TG	TG	66.9	1.1	66.4	1.2	68.0	66.2
RO	TG_{M2M}	TG	65.7	1.1	66.0	2.5	66.7	65.8
RO	TG_{DL}	TG	67.6	1.1	67.0	1.7	67.7	66.5

Table 8: Scores and standard deviations of $F1_m$, $F1_{pn}$, an accuracy (acc) and macro averaged recall (rec) scores for machine translation experiments. These results complement those presented in Table 3 of the main paper.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
"Limitations" section provided in the template

☑ A2. Did you discuss any potential risks of your work?
"Ethics Statement" section provided in the template

☑ A3. Do the abstract and introduction summarize the paper's main claims?
"Abstract" and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Abstract, Section 1, 2, 3, 4, 5, Limitations, Ethics Statement

☑ B1. Did you cite the creators of artifacts you used?
Section 1, 2, 3, 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Subsection 3.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2, Subsection 3.1

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Subsection 3.1, Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3, Section 4

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A, Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4, Appendix C

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, Section 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix E

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics Statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix D