Cross-Lingual Rumour Stance Classification: a First Study with BERT and Machine Translation

Carolina Scarton and **Yue Li**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, S1 4DP, Sheffield
{c.scarton, yli381}@sheffield.ac.uk

Abstract

Social media tend to be rife with rumours, which often have such high velocity and volume that fact-checkers struggle with debunking them with traditional methods. Prior research on English rumours has demonstrated that one can analyse the reactions (i.e. stance) expressed by social media users towards rumours, which ultimately enables automated flagging to journalists highly disputed rumours. This paper presents the first study of cross-lingual rumour stance classification. Through experiments with zero- and few-shot learning and in three languages (German, Danish and Russian), we show that models trained on English data can be used successfully for predicting stance in other languages. In the few-shot case, we also show that only few data points in the target language are needed to achieve the best results. In a multilingual setting, results for English are also further improved. Our results highlight the potential of multilingual BERT and machine translation for rumour analysis in languages where annotated data is scarce or not readily available.

1 Introduction

Social media are rife with rumours, which are fast-spreading, unverified pieces of information (Zubiaga et al., 2018). Journalists, however, are struggling to analyse misinformation in real time and at scale, thus motivating research into automatic rumour detection and analysis. A key rumour analysis task is *rumour stance classification* (RSC) (Li et al., 2019b; Dungs et al., 2018; Aker et al., 2019). RSC is useful, for instance, to flag rumours in a human-in-the-loop model, where fact-checkers could rely on crowd-based information, i.e. replies that support or deny a given rumour (Karmakharm et al., 2019). State-of-the-art automatic models for veracity prediction have also successfully used stance information in order to achieve best results. It is typically modelled as a four class problem, where posts replying to a rumour are classified as supporting; denying; questioning; or commenting on the rumour (Procter et al., 2013). In particular, the RumourEval 2017 (Derczynski et al., 2017) and 2019 (Gorrell et al., 2019) shared tasks demonstrated that RSC is a highly imbalanced problem, where the most informative classes, namely *support* and *deny*, are the minority classes.

Previous research on RSC, however, has focused predominantly on English, with the exception of Lozhnikov et al. (2018) for Russian and Lillie et al. (2019) for Danish. In the former, a small dataset (958 data points) composed of tweets and comments to media headlines is used to train RSC classifiers with word embeddings as features achieving 0.865 of macro-$F1$. In the latter, annotated Reddit posts in Danish (DAST) are used to train RSC classifiers using a varied feature set that includes word embeddings, part-of-speech, sentiment, and meta-data information. DAST has $3,007$ data points and the best model (SVM) achieves 0.421 of macro-$F1$.

This paper presents, to the best of our knowledge, the first study on cross-lingual RSC. We explore both multilingual BERT (MBERT) (Devlin et al., 2019) and machine translation (MT) approaches for zero-shot learning, and MBERT for few-shot learning as well as to train a full multilingual model. We make use of English, German, Russian and Danish RSC datasets, aiming to transfer the knowledge from English into other languages. Although our best results for Russian (macro$-F1 = 0.506$) and Danish (macro$-F1 = 0.352$) are not directly comparable to the performance of the respective language-specific models from prior work, we argue that cross-lingual RSC is a harder problem and that, given the small number of data points, previous work may be suf-

fering from overfitting. Moreover, the models trained in previous work require reliable Natural Language Processing tools (e.g. part-of-speech taggers, sentiment analysers) readily available in languages other than English, which is not usually the case. Nevertheless, ours is the first study to apply established multilingual models to perform cross-lingual RSC and thus enable RSC in low-resourced languages.

2 Related Work

In this section we discuss work related to RSC, focusing on the RumourEval datasets. RumourEval is a shared task, organised as part of SemEval in 2017 and 2019, comprising two subtasks: (A) stance classification and (B) veracity prediction. In A, the stance classification is formulated as a four-class classification problem, where replies to a social media post (source) can support, deny, query or provide a comment to the source. Subtask B consists of predicting the veracity of a rumour in social media, based on the text and/or metadata features. Successful systems in task B have also used the result of RSC as features (Li et al., 2019a,b). RSC (as formulated in the RumourEval datasets) is different from traditional stance classification tasks, since it also proposes the *query* class, useful in the rumour identification scenario. In addition, as pointed out by Scarton et al. (2020), in RSC, classes have different importance, with *support* and *deny* being the most interesting class for the task. This is particularly useful for human-in-the-loop application, where knowing if a reply is denying or supporting a post is more informative than if the reply is a comment.

Previous research in this area mainly focuses on the special characteristics of the datasets. Kochkina et al. (2017) employ Long Short Term Memory networks (LSTM) to capture the sequential nature of tweet threads. Yang et al. (2019) propose an inference chain-based system that utilise the information of the whole conversation. Task-specific features are also designed to boost the classifier performance (Aker et al., 2017; Bahuleyan and Vechtomova, 2017; Ghanem et al., 2019). Dealing with skewed distribution towards the *comment* class is another significant direction as most of the systems suffer the low performance over the minority classes, especially the *deny* class. Li and Scarton (2020) compare the performance of traditional imbalanced data treatments on the RumourEval datasets. They design a simple BERT-based model combining with threshold-moving, ranking first and second in RumourEval 2017 and 2019 respectively.

However, RumourEval datasets only consider English posts. To the best of our knowledge, Lozhnikov et al. (2018) (for Russian) and Lillie et al. (2019) (for Danish) are the only previous work tackling RSC for languages other than English. For both, datasets are created following the same annotation scheme as RumourEval, i.e. replies to comments are annotated in one of the four classes (for more details about these datasets see Section 3.1). In (Lozhnikov et al., 2018), feature-based classifiers are trained, using pre-trained word embeddings for Russian. They achieve an impressive 0.865 of macro-$F1$, with great performance for *support* and *deny* classes. In the work for Danish, they experiment with a LSTM classifier and several feature-based models, using the DAST dataset. Different feature types were explored, including textual information, sentiment, bag-of-words, part-of-speech, word frequency and information from the Reddit metadata. Their best model, achieving 0.421 of macro-$F1$, is an SVM trained with hyperparameter optimisation and feature selection (e.g. Reddit-based features are not included in this model). (Lillie et al., 2019) is also the first work to present cross-lingual veracity prediction. Veracity classifiers are trained relying on language independent information using the PHEME dataset (Zubiaga et al., 2016) for training and DAST for testing and vice-versa. Results suggest that cross-lingual models are comparable to monolingual models.

Although these two previous work represent an advance in RSC for languages other than English, the processing of collecting monolingual annotated data is expensive and time-consuming to be feasible for all languages. In addition, they assume that NLP resources, such as pre-trained word embeddings, part-of-speech taggers and sentiment analysis models, are readily available in the language under study, which is not a reality for most languages other than English. Therefore, it is important to explore approaches that enable low-resourced languages to benefit from the relatively large amount of English training data. Although cross-lingual approaches have been investigated for various NLP problems (Stappen et al., 2020; Chidambaram et al., 2019; Eriguchi et al., 2018),

	Support	Deny	Question	Comment
EN(training)	841 (19.8%)	333 (7.8%)	330 (7.8%)	2,734 (64.5%)
EN(test)	94 (9.0%)	71 (6.8%)	106 (10.1%)	778 (74.1%)
DE (test only)	48 (17.0%)	13 (4.6%)	18 (6.4%)	203 (72.0%)
DA(training)	184 (8.2%)	232 (10.4%)	61 (2.7%)	1756 (78.6%)
DA(test)	89 (11.5%)	68 (8.8%)	20 (2.6%)	597 (77.1%)
RU(training)	35 (5.0%)	36 (5.2%)	139 (20.1%)	481 (69.6%)
RU(test)	23 (8.6%)	10 (3.7%)	53 (19.9%)	181 (67.8%)

Table 1: Data distribution of classes in each dataset (values in parenthesis are the percentages of each class).

to the best of our knowledge, there is no research on cross-lingual RSC task.

3 Experimental Settings

3.1 Datasets

English The English model is trained on the RumourEval 2017 (RE2017) dataset (Derczynski et al., 2017) which has $4,238$ source-reply tweet pairs from eight different events in the training set: the Ferguson unrest, the shooting at Charlie Hebdo, the hostage situation in Sydney, the Germanwings plane crash, the Ottawa shooting, a rumour about a coup in Russian, a rumour that Prince was doing a surprise show in Toronto, and a rumour that Footballer Michael Essien had contracted Ebola. The test set has $1,049$ tweet pairs from ten events (the same eight events in the training data plus: a rumour that Hillary Clinton was diagnosed with pneumonia during the 2016 US elections and rumour that Youtuber Marina Joyce had been kidnapped).

German The German data (Zubiaga et al., 2016) has 282 tweet pairs from three different events: the Germanwings plane crash, a rumour about a coup in Russian, and a rumour about the Gurlitt collection.

Danish For Danish, we use the DAST dataset with $3,007$ source-reply Reddit pairs (Lillie et al., 2019). It encompasses posts from 11 rumourous events: 5G, Donald Trump, HPV vaccine, ISIS, *Kost* (diet), MeToo movement, *Overvågning* (surveillance), Peter Madsen, *Politik* (politics), *Togstrejke* (train strike), and *Ulve i DK* (wolves in Denmark).

Russian For Russian we use a dataset with source-reply tweet pairs concatenated with claim-reply pairs of Meduza[1] and Russian Today.[2] It has 958 pairs divided into 17 threads covering different topics (Lozhnikov et al., 2018).[3]

For monolingual and few-shot learning experiments, we divide the Danish and Russian datasets into training and test sets. For Danish, eight events are used for training (ISIS, *Kost*, MeToo, *Overvågning*, Peter Madsen, *Politik*, *Togstrejke*, and *Ulve i DK*) and three for testing (5G, HPV vaccine, and Donald Trump). For Russian, 14 topics are used for training and three for testing. Dividing the training and test sets using events/topics is expected to minimise the chances of overfitting, since, at training time, the models will not see the events that appear in the test set. The German dataset is rather small, with only one of the three events having data points in all classes, which makes it unsuitable for data splitting. Therefore, this dataset is only used as a test set in the zero-shot experiments. Table 1 shows the class distributions for each dataset.

3.2 Models

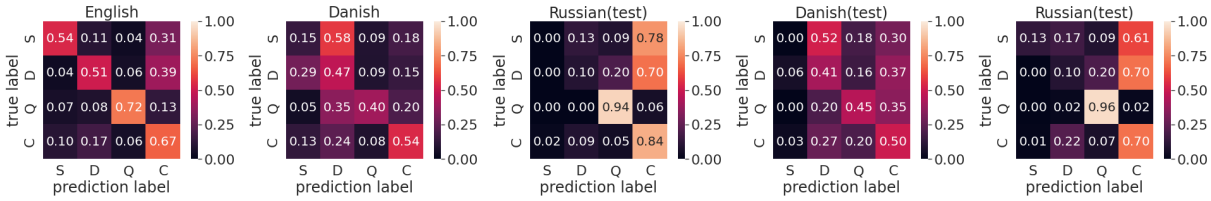
Settings BERT models are fine-tuned for three epochs with a batch size of 16, 12 transformer layers, hidden unit size of 768, 12 attention heads, and 110M parameters using the `ktrain` toolkit (Maiya, 2020). We apply the 1 *cycle policy* (Smith, 2018) for training and search the optimal learning rate among $5e^{-5}$, $3e^{-5}$, $1e^{-5}$, and $1e^{-4}$. For dealing with data imbalance, we follow (Li and Scarton, 2020) and apply threshold moving (TM) (Maloof, 2003; Sheng and Ling, 2006), where the classifier is trained with the imbalanced data, but the decision threshold that transforms the output probability into class labels is changed. We set the threshold according to the class proportions based on two assumptions: (1) the class proportion in the test set is similar to that of the training set; and (2) the prior of a class is equivalent to its pro-

[1] https://meduza.io/en
[2] https://www.rt.com

[3] More details about the topics are available at (Lozhnikov et al., 2018).

		macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
SOTA RE2017	EN(test)	0.452	0.363	0.296
MBERT_EN	EN(test)	**0.528**	**0.602**	**0.487**
MBERT_DA	DA(test)	**0.300**	**0.350**	**0.251**
MBERT_RU	RU(test)	0.442	0.000	0.211
MBERT_MTDA	DA(test)	0.228	0.00	0.166
MBERT_MTRU	RU(test)	**0.467**	**0.306**	**0.259**

Table 2: Results for the monolingual MBERT models (best results are shown in bold).



(a) MBERT_EN (b) MBERT_DA (c) MBERT_RU (d) MBERT_MTDA (e) MBERT_MTRU

Figure 1: Confusion matrices for monolingual models.

portion in the training set (Collell et al., 2018).

Multilingual BERT (MBERT) use the pre-trained *BERT-base-multilingual-cased* model.[4] The hypothesis of using a MBERT model trained only on English data (MBERT_EN) for other languages is that it would be capable of performing zero-shot RSC, similar to its success in other NLP tasks (Pires et al., 2019). We also experiment with models trained only on Danish (MBERT_DA) or Russian (MBERT_RU) training data. *Few-shot* learning is also explored, where MBERT_EN is further fine-tuned using the training data in Danish (MBERT_ENDA) or Russian (MBERT_ENRU). We aim to check whether monolingual data, even just a few data points, can help to improve the performance of SRC in Danish or Russian. Finally, we also propose a full multilingual model, where the English, Danish and Russian training sets are combined and used for training the model.

Machine Translation (MT) is used to translate the Russian, Danish and German data into English, so the English-only models can be applied. We use Google Translate[5] for producing the automatic translations and MBERT_EN to classify the translated text (MT+MBERT_EN model). We also use MT to translate the English training data into Russian or Danish, and fine-tune MBERT monolingual models (MBERT_MTDA and MBERT_MTRU models for Danish and Russian, respectively).

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://translate.google.co.uk

3.3 Evaluation

Scarton et al. (2020) show that the evaluation metrics used in RumourEval in 2017 (accuracy) and 2019 (macro-$F1$) are not robust for this four-class imbalanced classification task. They suggest the use of two alternative metrics: geometric mean of recall (GMR) and $wF2$. GMR heavily penalises models that underperform on minority classes, being an useful metric for imbalanced classification tasks. $wF2$ is a weighted version of macro-$F2$ that gives more importance to recall than precision and also assigns higher weights for the most important RSC classes, i.e. *support* and *deny*. Therefore, in this paper, besides reporting macro-$F1$ for comparison with previous work, we also report $wF2$ and GMR.[6]

4 Cross-lingual Rumour Stance Classification

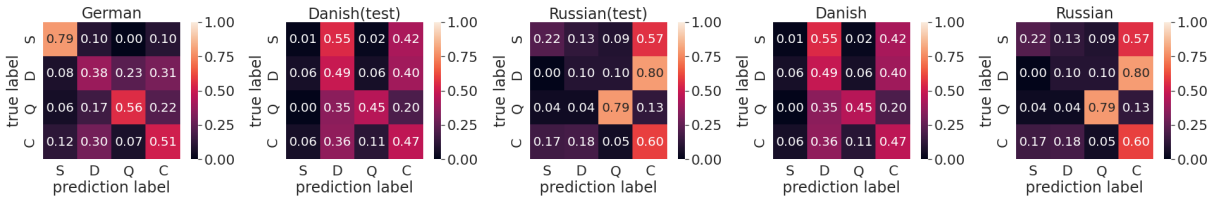
4.1 Monolingual models

Aiming to assess the effectiveness of zero- and few-shot models, we devise monolingual models for Danish and Russian, either using our pre-defined training sets (MBERT_DA and MBERT_RU) or the machine translated training sets (MBERT_MTDA and MBERT_MTRU). Results shown in Table 2 also include values for MBERT_EN in the RE2017 test set and for the best model in the RE2017 shared task (Best

[6] For $wF2$, we use the same weights as Scarton et al. (2020), i.e $w_{deny} = w_{support} = 0.40$, $w_{query} = 0.25$ and $w_{comment} = 0.05$

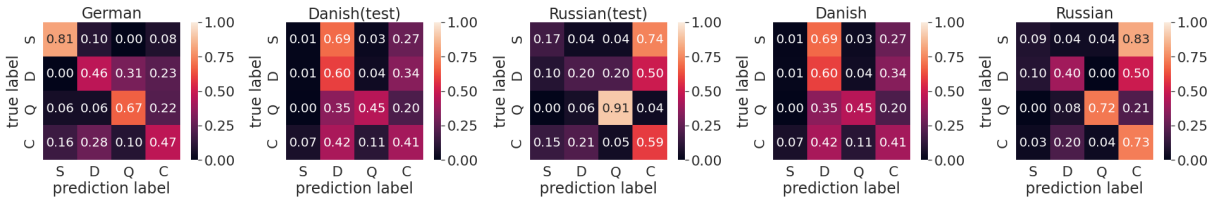
	MBERT_EN			MT+MBERT_EN		
	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
DE	**0.470**	0.542	0.480	0.464	**0.585**	**0.505**
DA(full)	**0.259**	0.221	0.201	0.248	**0.228**	**0.219**
DA(test)	**0.241**	0.184	0.187	0.234	**0.188**	**0.200**
RU(full)	**0.419**	**0.377**	**0.278**	0.406	0.360	0.260
RU(test)	0.420	0.319	0.252	**0.437**	**0.368**	**0.275**

Table 3: Results for zero-shot learning using MBERT_EN model or MT+MBERT_EN. Best values are in bold.



(a) (b) (c) (d) (e)

Figure 2: Confusion matrices for zero-shot learning using MBERT_EN.



(a) (b) (c) (d) (e)

Figure 3: Confusion matrices for zero-shot learning using MT+MBERT_EN.

RE2017), showing that MBERT_EN would have ranked first in this shared task.[7] Interestingly, MBERT_MTRU performs significantly better than MBERT_RU, which may be explained by the small size of the Russian training set. Conversely, MBERT_DA performs best, probably due to more in-domain data available.

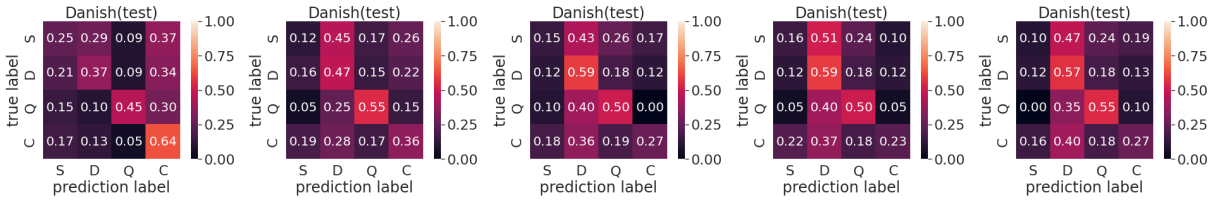
Figure 1 shows the confusion matrices for the monolingual models. MBERT_RU has $GMR = 0$ because it fails to predict all *support* instances. MBERT_DA also underperforms for the *support* class and has a bias towards *denies* (mainly by predicting *supports* as *denies*). MBERT_DA is the best for Danish, since it predicts 15% of *support* and 47% of *deny* correctly, versus 0% and 41% for *support* and *deny*, respectively, for MBERT_MTDA.

4.2 Zero-Shot Rumour Stance Classification

In the first zero-shot experiment, MBERT_EN model is used for RSC in other languages. We then compare this model with a pipelined approach using MT+MBERT_EN, where the data in Danish, Russian or German are machine translated into English and classified using MBERT_EN. Table 3 shows the results of evaluation in the Danish and Russian full and test sets and the German set, whilst Figure 2 and Figure 3 show the confusion matrices for models MBERT_EN and MT+MBERT_EN, respectively. Results for German are particularly good, being comparable to the results for English (Table 2). One reason for this high performance is that the German test set includes tweets about rumours that appear in the English training set.

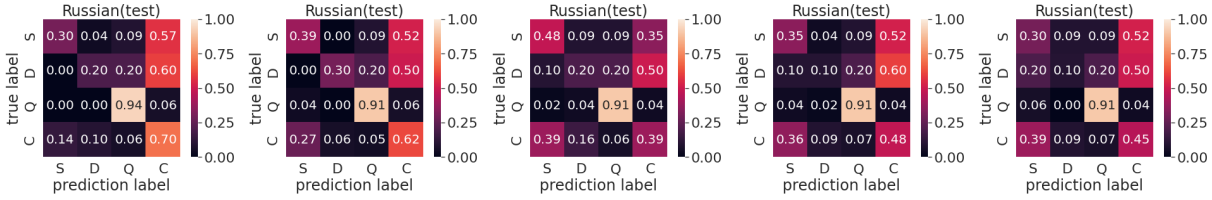
For Danish, the best GMR and $wF2$ are achieved with the pipelined MT+MBERT_EN model (for both full and test), however, these results are worse than the monolingual model (Table 2). The main issue is with the misclassification of *supports* (-0.14 of class accuracy in comparison to MBERT_DA in Figure 1b). Data characteristics may justify this low performance: while the English training data is composed of tweets, the Danish data has Reddit posts, which are consid-

[7] The best model for RE2017 according to the reported metrics is NileTMRG (Enayet and El-Beltagy, 2017). This differs from the winner of the task (Kochkina et al., 2017), which shows low scores for all metrics (Scarton et al., 2020).



(a) Full training set (b) $n = 244$ (c) $n = 100$ (d) $n = 60$ (e) $n = 20$

Figure 4: Confusion matrices for few-shot learning using MBERT_EN model as starting point for Danish (MBERT_ENDA).



(a) Full training set (b) $n = 140$ (c) $n = 100$ (d) $n = 60$ (e) $n = 20$

Figure 5: Confusion matrices for few-shot learning using MBERT_EN model as starting point for Russian (MBERT_ENRU).

erably longer and use different argumentation patterns.[8] For Russian, MT+MBERT_EN also shows best results for the test set with improvements due to better performance in the *supports* class (+0.17 in comparison to MBERT_RU in Figure 1c). On the other hand, the full set in Russian achieves best results when MTBERT_EN is applied.

For German (Figures 2a and 3a) and Danish (full: 2d and 3d; and test: 2b and 3b), the best results in GMR and $wF2$ for MT+BERT_EN is explained because this model outperforms BERT_EN for classes *support* and *deny*. For RU(full), MBERT_EN (Figure 2e) is significantly better at predicting *denies* than MT+MBERT_EN (Figure 3e). On the other hand, for RU(test) MT+BERT_EN (Figure 3c) shows significantly better results for *deny* and *query* classes than BERT_EN (Figure 2c).

4.3 Few-shot Rumour Stance Classification

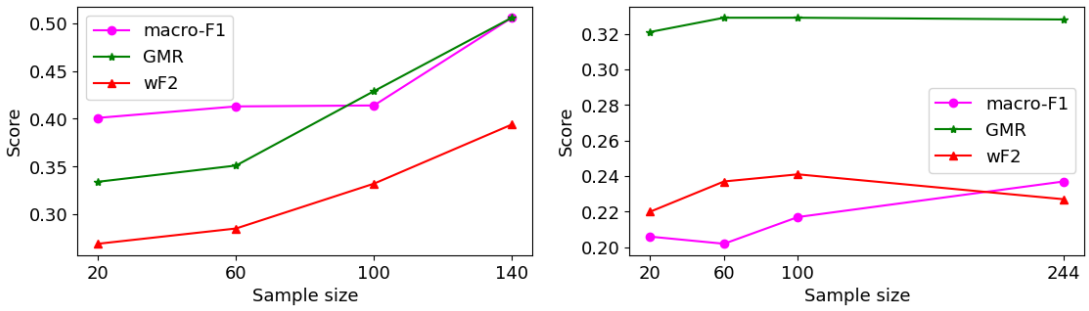
For few-shot learning, we use MBERT_EN as the starting point and continue fine-tuning it with the target language training data, i.e. either DA(training) or RU(training). When monolingual data is available for training, the hypothesis is that MBERT models would benefit from the pre-training on a larger dataset (English) and specialise their performance using target language data. Table 4 shows the performance of models trained in this setting: MBERT−ENDA for Danish (the confusion matrix is show in Figure 4a) and MBERT−ENRU for Russian (the confusion matrix is show in Figure 5a). Results for Russian show a significant increase in performance over the monolingual and zero-shot models, specially in terms of GMR and $wF2$. This happens mainly due to improvements in the accuracy of *supports* (+0.30) and *denies* (+0.10). Few-shot learning also improves the results for Danish, mainly because the MBERT_ENDA model better handles all classes (specially *support*, improving +0.10 points), without biasing towards *denies*.

	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
DA(test)	**0.352**	**0.401**	**0.295**
RU(test)	0.501	0.448	0.349
DA(balanced)	0.237	0.328	0.227
RU(balanced)	**0.506**	**0.506**	**0.394**

Table 4: Few-shot learning using MBERT_EN model as starting point (best results are shown in bold).

Balanced data re-sampling We under-sampled the Russian and Danish training sets, so that all classes have the same number of data points. For Russian, since the class with fewest examples (*support*) has 35 instances, 140 is the size of this balanced training set. For Danish, the smallest class is *query* with 61 examples, so the balanced set has 244 data points. Results for models trained

[8] We have also experimented with an MBERT_EN model trained on RumourEval 2019 data that contains Reddit posts. Results for Danish did not improve, probably due to the size of the English Reddit sample (only 16.9% of the training data).



(a) MBERT_ENRU (b) MBERT_ENDA

Figure 6: Performance of: (a) MBERT_ENRU varying the sample sizes of the Russian training set and MBERT_ENDA varying the sample sizes of the Danish training set.

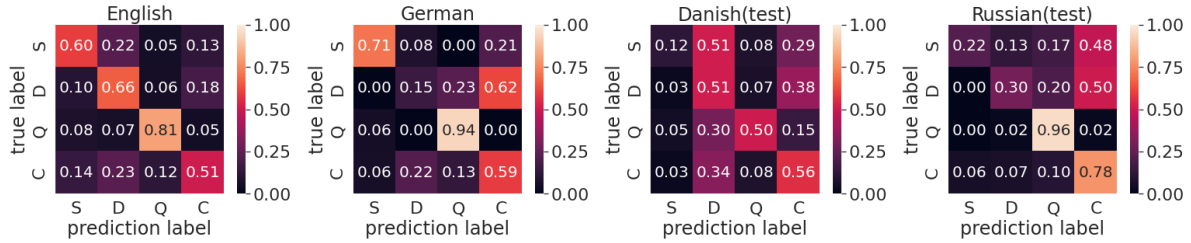


Figure 7: Confusion matrices for the full multilingual model.

on these balanced sets are shown in the bottom part of Table 4: DA(balanced) (see Figure 4b for the confusion matrix) and RU(balanced) (see Figure 5b for the confusion matrix). For Russian, this approach outperforms the use of the entire training set, with further improvements in the *support* (+0.09) and *deny* (0.10) classes. For Danish, the balanced approach is not better than using the entire training set.

Monolingual data sample size Aiming to assess the amount of monolingual data needed to outperform zero-shot learning, we also decrease the size of the samples gradually, starting from 35 (or 61) and stopping at 5 data points per class. For Russian (Figure 6a), 100 data points of balanced training data is enough to outperform zero-shot approaches ($GMR = 0.429$ and $wF2 = 0.332$). For Danish (Figure 6b), 20 data points is enough to improve over zero-shot learning, with $GMR = 0.303$ and $wF2 = 0.220$. The confusion matrices for this experiment are also shown: Figures 4b to 4e for Danish and Figures 5b to 5e for Russian. We observe that for Danish, the best performance for *support* class is achieved when the full dataset is used (Figure 4a), whilst the best performance for *denies* is reached when the samples size (n) is 60 or 100. In particular, there is a significant drop in the performance of *denies* when the data sample is increased to $n = 244$, which justifies the decrease in $wF2$ show in Figure 6b. The main issue with the balanced Danish models is the bias towards the *deny* class, which is minimised when the full Danish traning set is used. For Russian, the best performance for *support* is at $n = 100$, whilst the *deny* is more accurately predicted when $n = 140$.

4.4 Full multilingual model

To build a *full* multilingual model, we fine-tune MBERT with all training data in all languages. We aim to assess whether joint training improves performance in the individual languages. Table 5 shows the results for this experiment and Figure 7 shows the confusion matrices.

	macro-$F1\uparrow$	$GMR\uparrow$	$wF2\uparrow$
EN	0.484	0.635	0.509
DA(test)	0.323	0.366	0.263
RU(test)	0.524	0.470	0.368
DE (zero-shot)	0.502	0.497	0.462

Table 5: Results for a MBERT fine-tuned with all training data for all languages.

Significant improvements are shown for English in terms of GMR and $wF2$, thanks to significant improvements in accuracy on the *support* (+0.06), *deny* (+0.15) and *query* (+0.09) classes. Even

though the macro-$F1$ for English here is worse than the monolingual model (Table 2), our multilingual model still outperforms the state-of-the-art for the RE2017 shared task. Results for Danish and Russian are better than zero-shot models, although worse than few-shot models. In few-shot learning, the models get more specialised in the target language, while in the multilingual setting the variety of data may harm the prediction for languages with fewer data points. For German, this is also a zero-shot setting, since we do not have training data for this language. Results in terms of GMR and $wF2$ are worse in this multilingual setting than in our zero-shot experiment (Table 3), mainly because the performance of *support* is significantly harmed. We hypothesise that the variety of data introduced by Danish and Russian can be harming the performance for German, that is a very similar set to the English training data.

5 Conclusions

To the best of our knowledge, this is the first paper to produce a detailed comparison of cross-lingual RSC on four languages (English, German, Danish, and Russian) and across different types of posts (tweets, Reddit posts, and comments to media headlines). The results of our zero-shot learning experiments show that both MT- and MBERT-based RSC can be useful for low-resourced languages, where no data is available for training.

Few-shot learning shows the best performance for both Danish and Russian, outperforming zero-shot models with just a few data points in the target language. Therefore, monolingual data can be useful for improving models, but only a few data points are actually needed (in our experiments, models outperforming the zero-shot experiments were achieved with 100 data points for Russian and 20 for Danish). A full multilingual model improved the performance for English, showing that data in other languages may also be helpful for high-resource languages.

We argue that cross-lingual RSC can also enable the analysis of trending rumours, that may have replies in multiple languages. In particular, MBERT-based approaches can also be useful for robustly model code-switching, where a single reply contains words in multiple languages. Future work include further experiments with more languages (given the availability of data) and the use of cross-lingual RSC for supporting the task of veracity prediction.

Acknowledgements

This work was funded by the WeVerify project (EU H2020, grant agreement: 825297).

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple open stance classification for rumour analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria. INCOMA Ltd.

Ahmet Aker, Alfred Sliwa, Fahim Dalvi, and Kalina Bontcheva. 2019. Rumour verification through recurring information and an inner-attention mechanism. *Online Social Networks and Media*, 13:100045.

Hareesh Bahuleyan and Olga Vechtomova. 2017. UWaterloo at SemEval-2017 task 8: Detecting stance towards rumours with topic independent features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 461–464, Vancouver, Canada. Association for Computational Linguistics.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Guillem Collell, Drazen Prelec, and Kaustubh R. Patil. 2018. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, 275:330–340.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Omar Enayet and Samhaa R. El-Beltagy. 2017. NileTMRG at SemEval-2017 task 8: Determining rumour and veracity support for rumours on Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474, Vancouver, Canada. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Bilal Ghanem, Alessandra Teresa Cignarella, Cristina Bosco, Paolo Rosso, and Francisco Manuel Rangel Pardo. 2019. UPV-28-UNITO at SemEval-2019 task 7: Exploiting post's nesting and syntax information for rumor stance classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1125–1131, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Twin Karmakharm, Nikolaos Aletras, and Kalina Bontcheva. 2019. Journalist-in-the-loop: Continuous learning as a service for rumour analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 115–120, Hong Kong, China. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019a. eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019b. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy. Association for Computational Linguistics.

Yue Li and Carolina Scarton. 2020. Revisiting rumour stance classification: Dealing with imbalanced data. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 38–44, Barcelona, Spain (Online). Association for Computational Linguistics.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Nikita Lozhnikov, Derczynski Leon, and Mazzara Manuel. 2018. Stance Prediction for Russian: Data and Analysis. In *Proceedings of 6th International Conference in Software Engineering for Defence Applications*, pages 176–186, Roma, Italy. Advances in Intelligent Systems and Computing, Springer, Cham.

Arun S. Maiya. 2020. ktrain: A Low-Code Library for Augmented Machine Learning. *arXiv preprint arXiv:2004.10703*.

Marcus A Maloof. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Proceedings of the ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, DC, USA.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214.

Carolina Scarton, Diego Silva, and Kalina Bontcheva. 2020. Measuring what counts: The case of rumour stance classification. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 925–932, Suzhou, China. Association for Computational Linguistics.

Victor S. Sheng and Charles X. Ling. 2006. Thresholding for making classifiers cost-sensitive. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, pages 476–481, Boston, Massachusetts. American Association for Artificial Intelligence.

Leslie N. Smith. 2018. *A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay*. US Naval Research Laboratory Technical Report 5510-026.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel. *arXiv preprint arXiv:2004.13850*.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.