Fine-tuning Neural Language Models for Multidimensional Opinion Mining of English-Maltese Social Data

Keith Cortis
ADAPT Centre,
Dublin City University,
Glasnevin, Dublin 9, Ireland
keith.cortis
@adaptcentre.ie

Kanishk Verma
ADAPT Centre,
Dublin City University,
Glasnevin, Dublin 9, Ireland
kanishk.verma
@adaptcentre.ie

Brian Davis
ADAPT Centre,
Dublin City University,
Glasnevin, Dublin 9, Ireland
brian.davis
@adaptcentre.ie

Abstract

This paper presents multidimensional Social Opinion Mining on user-generated content gathered from newswires and social networking services in three different languages: English —a high-resourced language, Maltese —a low-resourced language, and Maltese-English —a code-switched language. Multiple fine-tuned neural classification language models which cater for the i) English, Maltese and Maltese-English languages as well as ii) five different social opinion dimensions, namely subjectivity, sentiment polarity, emotion, irony and sarcasm, are presented. Results per classification model for each social opinion dimension are discussed.

1 Introduction

Social Opinion Mining on data obtained from social sources is an evolving research domain tasked with the identification of several opinion dimensions, such as *subjectivity*, *sentiment polarity*, *emotion*, *irony* and *sarcasm*, from noisy user-generated social data spread across heterogeneous sources (Cortis and Davis, 2021b). Currently, Social Opinion Mining is used in several real-world scenarios, namely chatbots (Androutsopoulou et al., 2019), adaptive customer online service based on identified customer sentiment and emotion (Yadollahi et al., 2017), tracking of overall customer satisfaction for a product or service (Zhao et al., 2019), and detection of changes in customer opinion towards a brand, product or service (Geetha et al., 2017).

This paper presents multidimensional Social Opinion Mining on user-generated content gathered from newswires and social networking services in three different languages: **English** – a high-resourced language, **Maltese**[1] – a low-resourced language, and **Maltese-English** – a code-switched language. Our aim is use these initial results to improve cross-lingual performance of English-Maltese neural language models. Research applications of the developed classification models include opinion summarisation and fine-grained opinionated search of each dimension. This work is in line with Malta's Strategy and Vision for Artificial Intelligence (Parliamentary Secretariat for Financial Services and Innovation, 2019), with current investment being made in the development of Maltese language resources and tools to counter the threat of "digital extinction" for the Maltese language, which has low technological support available in comparison with other European languages (Rosner et al., 2012).

We leverage a novel multidimensional and multilingual social opinion dataset in the socio-economic domain, specifically Malta's annual Government Budget, which comprises social data from the 2018, 2019 and 2020 budgets to fine-tune pre-trained neural language models for benchmarking purposes.

2 Related Work

Nguyen et al. (Nguyen et al., 2020) developed the first large-scale pre-trained language model BERTweet for English tweets, which outperforms its baselines. Experiments were conducted on three NLP tasks, namely Part-of-Speech tagging, Named Entity Recognition and text classification, namely sentiment analysis and irony detection. For the latter task, the authors used the 3-class sentiment analysis dataset from SemEval-2017 Task 4A (Rosenthal et al., 2017) and the 2-class irony detection dataset from the SemEval-2018 Task 3A (Van Hee et al., 2018). The authors in (Croce et al., 2020) propose GAN-BERT which extends the fine-tuning of architectures similar to Bidirectional Encoder Representations from Transformers (BERT) (Devlin

[1] Maltese (Malti) is a Semitic language written in Latin script and is Malta's national language

Proceedings of Recent Advances in Natural Language Processing, pages 309–314
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_036

et al., 2018), using unlabelled data in a generative adversarial setting. Experimental results show that around 50-100 annotated examples can still produce good performance in sentence classification tasks. Results are confirmed for sentiment analysis over the SST-5 dataset (Socher et al., 2013) containing 5-class sentiment polarity categories. Babanejad et al. (Babanejad et al., 2020) propose two novel deep neural network models for sarcasm detection by including affective and contextual features in the extended BERT architecture.

Certain studies focused on low-resourced languages, with (Fei and Li, 2020) investigating cross-lingual sentiment classification where the low-resource language does not have any labels or parallel corpus, (Grießhaber et al., 2020) exploring the reduction of trainable model parameters for fine-tuning a model with a small amount of data, (Koto et al., 2020) releasing a new pre-trained language model for Indonesian which was evaluated on several tasks such as sentiment analysis, and (Yimam et al., 2020) using RoBERTa (Liu et al., 2019)–a replication of BERT developed by Facebook– for exploring Amharic sentiment analysis from social media text.

Demszky et al. (Demszky et al., 2020) conduct transfer learning experiments on existing emotion benchmarks to show that the GoEmotions dataset of fine-grained emotions generalises across domains and taxonomies. The authors demonstrate that if little target domain labelled data is available, this dataset can be used as a baseline for emotion understanding. Similarly, the XED multilingual dataset for emotion detection catering for a total of 32 languages has been evaluated using language-specific BERT models (Öhman et al., 2020). Lastly, (Makarenkov and Rokach, 2020) explore several off-the-shelf BERT models, where they show that the complexity and computational cost of BERT does not provide a guarantee for an improved predictive performance for classification tasks. This is especially relevant in cases where small domain-specific datasets are used, which datasets are also imbalanced due to the minority class being under-represented.

3 Dataset

The dataset of multidimensional and multilingual social opinions for Malta's Annual Government Budget[2] (Cortis and Davis, 2021a) is used for the work carried out in this paper. This dataset contains 6,387 online posts for the 2018, 2019, and 2020 budgets, which user-generated content was collected from newswires and social networking services. In terms of languages, the majority of the online posts were in English (74.09%), Maltese or Maltese-English (24.99%). Each online post is annotated for the following five social opinion dimensions: *subjectivity*, *sentiment polarity*, *emotion*, *sarcasm* and *irony*. Table 1 presents the overall class distribution of online posts for each social opinion dimension and the language annotation. Statistics are provided for the entire dataset (columns 2 and 3), the subset of online posts in English (columns 4 and 5) and subset of online posts in Maltese and Maltese-English (columns 6 and 7).

4 Experiments

All experiments have been carried out on Google Colaboratory[3] using a Tesla K80/Tesla T4/Tesla P100-PCIE-16GB Graphics Processing Unit (GPU).

The baseline models experiments have been carried out in the Python programming language using Jupyter Notebook[4] on a machine with an Intel(R) Core(TM) i7-8550U CPU @ 1.80Hz 1.99 GHz processor and 8.00 GB (7.88 GB usable) installed memory (RAM).

4.1 Setup

We present multiple classification language models which cater for the English, Maltese and Maltese-English languages as well as **five** different social opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony* and *sarcasm*. We train models using state-of-the-art deep neural network models for each of the five opinion dimensions using the Transformer model architecture introduced by Vaswani et al. (Vaswani et al., 2017), which is based on attention mechanisms and is designed to handle sequential data, such as natural language, for NLP tasks like sentiment analysis and text summarisation.

4.2 Handling Imbalanced Data

As reflected in Table 1, the dataset we use is imbalanced. There are several re-sampling techniques (Cateni et al., 2014; More, 2016) for treating the problem of an imbalanced dataset. For our initial

[2] https://doi.org/10.5281/zenodo.4650232

[3] https://colab.research.google.com/

[4] https://jupyter.org/

Dataset	All		English		Maltese-English and Maltese	
	Count	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**
Subjectivity						
Subjective (1)	2591	40.57%	1713	36.20%	852	53.38%
Objective (0)	3796	59.43%	3019	63.80%	744	46.62%
Sentiment Polarity						
Negative (0)	1232	19.29%	775	16.38%	441	27.63%
Neutral (1)	1605	25.13%	1355	28.63%	219	13.72%
Positive (2)	3550	55.58%	2602	54.99%	936	58.65%
Emotion						
Joy (0)	2636	41.27%	1976	41.76%	648	40.60%
Trust (1)	363	5.68%	219	4.63%	144	9.02%
Fear (2)	72	1.13%	61	1.29%	11	0.69%
Surprise (3)	177	2.77%	116	2.45%	60	3.76%
Sadness (4)	245	3.84%	176	3.72%	67	4.20%
Disgust (5)	498	7.80%	275	5.81%	216	13.53%
Anger (6)	369	5.78%	238	5.03%	127	7.96%
Anticipation (7)	2027	31.74%	1671	35.31%	323	20.24%
Sarcasm						
Sarcastic (1)	177	2.77%	101	2.13%	74	4.64%
Not Sarcastic (0)	6210	97.23%	4631	97.87%	1522	95.36%
Irony						
Ironic (1)	329	5.15%	189	3.99%	136	8.52%
Not Ironic (0)	6058	94.85%	4543	96.01%	1460	91.48%
Language						
English (0)	4732	74.09%	4732	100%		
Maltese (1)	299	4.68%			299	18.73%
Maltese-English (2)	1297	20.31%			1297	81.27%
Other (3)	59	0.92%				

Table 1: Class distribution for each annotation per dataset

experiments, we do not address the imbalance or explore whether it influences our classification tasks and if so, which ones. The dataset was divided in a training set of 70%, validation set of 20% and a test set of 10%. The scikit-learn[5] train_test_split function is used to split the sets in a random state.

4.3 Models

The following state-of-the-art deep neural network models have been fine-tuned for *subjectivity* (binary), *sentiment polarity* (multi-class), *emotion* (multi-class), *sarcasm* (binary) and irony (binary) classification:

- **BERT** (Devlin et al., 2018): A pre-trained model on BookCorpus and English Wikipedia. The BERT-Base uncased, 12-layer, 768-hidden, 12-heads, 110M parameters model is used.
- **DistilBert** (Sanh et al., 2019): A distilled version of the BERT model which is smaller and faster than BERT and is pre-trained on the data. The uncased model which has 40% less parameters than BERT-Base uncased is used.
- **BERTweet** (Nguyen et al., 2020): A large-scale language model pre-trained for English tweets based on the RoBERTa (Liu et al., 2019) pre-training procedure using the same model configuration as BERT-Base. Both *bertweet-base* (base) and *bertweet-covid19-base-uncased* (covid-19) models with 135M parameters each are used. The former model is trained on 845M English cased tweets, whereas the latter model is trained on 23M COVID-19 English uncased tweets.

The experiments are carried out using the Hugging Face (Wolf et al., 2019) state-of-the-art Transformer library for Pytorch and TensorFlow 2.0[6]. This tool provides general-purpose architectures, such as BERT, RoBERTa and DistilBert for NLP tasks, such as sentiment analysis, where over 32+ pre-trained models are available in 100+ languages.

The following hyperparameters are used:

- Optimiser and learning rate scheduler: batch size - 32, Adam (Kingma and Ba, 2014) learning rate - 2e-5, number of epochs - 4, epsilon parameter - 1e-8;
- Method of choosing values and criterion used: Manual tuning based on training and validation loss, learning rates of 5e-5, 3e-5, 2e-5

[5] https://scikit-learn.org/

[6] https://huggingface.co/transformers/

and maximum sentence length of 96, 128 and 256 tokens;

- Fine-tuning classification layer: Rectified Linear Unit (ReLU).

5 Results and Discussion

Results per classification model for each social opinion dimension are presented in Table 2 and further discussed below. Three evaluation metrics are used to measure the classification performance of the fine-tuned models:

- **F1 score weighted**: F1 score is the weighted average of precision and recall. The weighted score calculates the F1 score for each label with their average being weighted by support, that is, the number of true instances for each label. This metric caters for label imbalance.
- **Area Under the Curve Receiver Operating Characteristics (AUC ROC)**: Score shows the model's true positive rate against the false positive rate and can help you identify how well (score of 1) a model can distinguish between classes[7].
- **Matthews correlation coefficient (MCC)**: measures quality of binary and multi-class classifications by taking into account true and false positives and negatives and provides a balanced measure for imbalanced classes.

The following is an overview of the results and some observations.

- **Subjectivity**: For the BERT and DistilBERT models, the training and validation loss converged in epoch 2, whereas both BERTweet models converged in epoch 3. The BERTweet covid19-base-uncased fine-tuned model produced the best performance overall.
- **Sentiment Polarity**: The fine-tuned BERT and DistilBERT models converged in epoch 3, whereas both BERTweet models converged in epoch 4. The BERTweet covid19-base-uncased fine-tuned model also produced the best performance overall.
- **Emotion**: The fine-tuned BERT and DistilBERT models converged in epoch 4, whereas both BERTweet models did not converge by epoch 4 albeit close. An additional experiment showed convergence in epoch 6. In terms of performance, both BERT and DistilBERT fared best overall.

	F1 Score	AUC ROC	MCC
Subjectivity			
BERT	0.93	**0.983**	0.864
DistilBERT	0.93	0.980	0.851
BERTweet (base)	0.93	0.970	0.857
BERTweet (covid19)	**0.94**	0.975	**0.887**
Sentiment Polarity			
BERT	0.85	1 - **0.945**	0.748
DistilBERT	0.83		0.710
BERTweet (base)	0.86	0 - **0.961**	0.772
BERTweet (covid19)	**0.87**	2 - **0.964**	**0.781**
Emotion			
BERT	**0.60**	3 - **0.935** 4 - **0.847** 5 - **0.914** 6 - **0.894**	0.495
DistilBERT	**0.60**	0 - **0.913** 1 - **0.821** 7 - **0.882**	0.484
BERTweet (base)	0.58	2 - **0.862**	0.478
BERTweet (covid19)	0.59		**0.501**
Sarcasm			
BERT	**0.96**	0.858	0.073
DistilBERT	**0.96**	**0.879**	**0.265**
BERTweet (base)	**0.96**	0.873	0
BERTweet (covid19)	**0.96**	0.792	0
Irony			
BERT	**0.93**	0.883	0.179
DistilBERT	**0.93**	**0.896**	**0.240**
BERTweet (base)	0.92	0.862	0
BERTweet (covid19)	0.92	0.887	0

Table 2: Results of all the pre-trained models

- **Sarcasm**: All fine-tuned models performed similarly in terms of F1 score, with DistilBERT performing best overall. The BERTweet covid19-base-uncased model did not converge in epoch 4 albeit close.
- **Irony**: DistilBERT produced the best results overall, which model converged in epoch 3.
- **Language**: It is interesting to see English-based fine-tuned models adapt to non-English text. This Maltese-English and Maltese subset amounts to only a quarter of the dataset (1596 online posts). Initial results obtained are promising for building language models that are capable of handling code-switched data, which is common practise in countries like Malta. More in-depth experiments and qualitative analysis shall be beneficial to measure the adaptability of the English-based fine-tuned models to code-switched languages, such as Maltese-English.
- **Domain**: A socio-economic dataset (domain specific) has been used, with only 16.75% of the data being off-topic. The results ob-

[7] For the sentiment polarity and emotion multi-class models we only display the maximum AUC ROC score for each respective class

tained in our preliminary work demonstrate that fine-tuning models to new domains is possible when using deep neural network models.

- The DistilBERT model took less time for training and validation for all five classifiers.
- Even though the dataset is imbalanced, the *subjectivity* and *sentiment polarity* models produced good results. However, certain resampling techniques shall help increase the performance of the *sarcasm* and *irony* fine-tuned models, which class distribution is very unbalanced as reflected by the MCC. The same also applies to the emotion model for certain classes, such as fear, surprise, sadness.
- Several researchers recommend only 2-4 epochs of training for fine-tuning BERT on a particular NLP task. However, certain multi-class classification tasks with a large number of classes such as the emotion 8-class classification fine-tuned model, might require more than 4 epochs when certain models such as BERTweet are fine-tuned.
- Given that the dataset used contains a mix of newswire comments and tweets, the maximum sentence length in the dataset used is 867. Therefore, more experiments should be carried out using a higher maximum sentence length than the 128 tokens used. However, the high computation power needed for training such deep learning models should be taken in consideration to reduce the carbon footprint in terms of finance and the environmental (Strubell et al., 2019).

6 Conclusions and Future Work

We have leveraged a novel multidimensional and multilingual social opinion dataset in the socio-economic domain to fine-tune neural language models targeting English-Maltese social data for different opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *sarcasm* and *irony*. Even though our results are a work-in-progress, we have been encouraged by Xia et al. (Xia et al., 2020) to provide multilingual benchmarks which can be further used, evaluated and adapted for low-resourced languages. Research applications for the developed classification models include opinion summarisation and fine-grained opinionated search of each dimension.

Acknowledgments

This work is funded by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant number 13/RC/2106).

References

Aggeliki Androutsopoulou, Nikos Karacapilidis, Euripidis Loukis, and Yannis Charalabidis. 2019. Transforming the communication between citizens and government through ai-guided chatbots. *Government Information Quarterly*, 36(2):358–367.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Silvia Cateni, Valentina Colla, and Marco Vannucci. 2014. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135:32–41.

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

Keith Cortis and Brian Davis. 2021b. Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, pages 1–93.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Digital Economy Parliamentary Secretariat for Financial Services and Office of the Prime Minister Innovation. 2019. Malta: The ultimate ai launchpad - a strategy and vision for artificial intelligence in malta 2030. https://malta.ai/wp-content/uploads/2019/10/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf. Date Published: 2019-10-02.

M Geetha, Pratap Singha, and Sumedha Sinha. 2017. Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management*, 61:43–54.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*.

Victor Makarenkov and Lior Rokach. 2020. Lessons learned from applying off-the-shelf bert: There is no silverbullet. *arXiv preprint arXiv:2009.07238*.

Ajinkya More. 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Mike Rosner, Jan Joachimsen, Georg Rehm, and Hans Uszkoreit. 2012. *The Maltese language in the digital age*. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Patrick Xia, Shijie Wu, and Benjamin Van Durme. 2020. Which *BERT? A survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online. Association for Computational Linguistics.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060.

Yabing Zhao, Xun Xu, and Mingshu Wang. 2019. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76:111–121.