SentiMalti: A Maltese Sentiment Analysis Dataset and Models

Ian Caruana*, **Matthew Vella***, **Fabio Zammit***, **Kurt Micallef, Claudia Borg**
Department of Artificial Intelligence, University of Malta
ian.caruana.23@um.edu.mt, matthew.g.vella.23@um.edu.mt,
fabio.zammit.23@um.edu.mt, kurt.micallef@um.edu.mt, claudia.borg@um.edu.mt

Abstract

We present SentiMalti, a new Maltese social media sentiment resource and accompanying baselines. We scrape user-generated content from YouTube, Reddit, and Facebook, then apply a Maltese-aware preprocessing pipeline (cleaning, personally identifiable information anonymisation, sentence splitting, and sentence-level language filtering) to retain Maltese sentences while tolerating realistic code-switching. The resulting crowdsourced dataset contains 2,327 sentences annotated for positive (39%), negative (31%), and neutral (30%) sentiment. We integrate prior Maltese datasets to create a combined benchmark of 3,772 instances. We evaluate fine-tuned encoder models (BERTu, Glot500) and few-shot prompting with instruction-tuned multilingual LLMs (Aya-101, Gemma 2 Instruct 9B). On the full test set, five-shot Aya-101 attains 68.65 macro-F1, closely followed by a fine-tuned BERTu at 68.36 macro-F1. Error analysis reveals complementary strengths: BERTu better separates polarised classes, while Aya-101 tends to over-predict the neutral class. We release the dataset splits, code, and a fine-tuned BERTu model to facilitate further work in Maltese NLP and sentiment analysis.

Keywords: sentiment analysis, crowdsourcing, low-resource, Maltese

1. Introduction

Sentiment Analysis can be considered a standard NLP task that has reached a high level of maturity, especially in high-resource languages. Yet, robust performance remains lacking in a low-resource scenario, as seen in the case of Maltese. Although there were previous efforts to compile Maltese sentiment analysis datasets, the two datasets are small in size and one is domain specific (Dingli and Sant, 2016; Cortis and Davis, 2019), leading to coverage gap and open questions about generalisation to contemporary social media discourse. These sources were used to fine-tune BERTu (Micallef et al., 2022) for sentiment analysis. However, they found that the size of the data affected the results, with sentiment analysis performing the worst of the attempted tasks.

This work aims to address this gap by constructing and publicly releasing a Maltese social media sentiment dataset together with newly fine-tuned models. To this end, we scrape comments from platforms such as YouTube, Reddit, and Facebook, focusing on popular Maltese channels and community fora to capture a broad spectrum of vernacular usage. The data was processed through a unified pipeline for cleaning, anonymisation, sentence splitting, and sentence-level language filtering, retaining Maltese sentences with a realistic tolerance for code-switching to English. These sentences were then annotated through a crowdsourcing exercise, resulting in a new dataset of 2,327 sentences, labelled as positive, negative, or neutral.

In this work, we also benchmark several language models using two approaches: (i) fine-tuning encoder-only models (BERTu (Micallef et al., 2022) and Glot500 (Imani et al., 2023)), and (ii) prompting instruction-tuned models (Aya-101 (Üstün et al., 2024) and Gemma 2 Instruct 9B Gemma Team et al. (2024)). Our evaluation shows that Aya-101 with five-shot prompting achieves the best performance with a macro-F1 of 68.65, followed closely by a fine-tuned BERTu, which obtains a macro-F1 score of 68.36.

Our key contributions are:

1. A new Maltese social media sentiment dataset and a pipeline for scraping, anonymising, and Maltese-aware filtering of sentences. We make the dataset publicly available to facilitate further research, improvements, and benchmarking in the area.[1]
2. A crowdsourcing website that allows the collection of majority-voting labelling for scraped sentences, prioritising under-annotated items and supporting quality control.
3. A comprehensive benchmark comparing fine-tuning versus few-shot prompting, including a shot-scaling and subset analysis of results. We also publicly release the fine-tuned BERTu model, which yields results comparable to larger prompted models.[2]

This work further diversifies sentiment analysis data and models for Maltese and provides an essential baseline for future advances in Maltese NLP.

*Equal contribution.

[1] https://huggingface.co/datasets/MLRS/SentiMalti

[2] https://huggingface.co/MLRS/BERTu_SentiMalti

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7927–7936
11–16 May 2026.

2. Related Work

2.1. Maltese Sentiment Data

In this section, we provide an overview of several data sources for Maltese Sentiment Analysis.

The first source from Dingli and Sant (2016) consists of 900 Maltese microblog comments from online news portals, referred to as **Microblogs** hereafter. The comments were manually annotated by three native speakers for sentiment polarity (positive, negative, and neutral). Their inter-annotator agreement was 51.22%, underscoring the difficulty of the task in an informal setting.

Subsequent work by Cortis and Davis (2019) focused on Malta's 2018 government budget and released the Social Opinion Gold Standard dataset, referred to as **Budget 2018** hereafter. This consisted of 555 comments collected from social media and news sources, annotated for sentiment (positive, negative, and neutral) and additional multidimensional labels (emotion and sarcasm). Annotation was performed by two raters, with an inter-annotator agreement of 60.15% for sentiment polarity (κ=0.3703), and a third expert consolidated cases of disagreement. This dataset included both Maltese and English material to reflect the bilingual nature of Maltese social media discourse.

A merged Maltese sentiment dataset, combining the above two datasets was made publicly available by Martínez-García et al. (2021). This dataset focused on binary sentiment labels (positive/negative), omitting the neutral class. The resulting dataset consisted of 271 positive samples (31.84%) and 580 negative samples (68.16%).

In Section 3, we present our data collection effort, which is significantly larger than these previous efforts. We also use these sources and combine them with our newly collected data to create a more balanced dataset in terms of label distribution and domain coverage.

2.2. Sentiment Analysis Models

Micallef and Borg (2025) benchmarked 55 large language models on several Maltese tasks, including Sentiment Analysis, using the dataset by Martínez-García et al. (2021). Out of all prompted models, Aya-101 (Üstün et al., 2024) was the best reported model across tasks (in both zero-shot and one-shot) and on Sentiment Analysis (in one-shot only), achieving macro-F1 scores of 78.1 and 86.5 for zero-shot and one-shot, respectively. In one-shot, Gemma 2 Instruct 9B also performs competitively with Aya-101. They also fine-tune BERTu (Micallef et al., 2022) on this data obtaining a macro-F1 score of 83.0. A fine-tuned mBERT, while being a competitive baseline, lags behind in performance, compared to the aforementioned models.

These results inform our choice of models in Section 4, although we also include Glot500 (Imani et al., 2023) since we observe more competitive performance on this task compared to mBERT (Devlin et al., 2019). Moreover, we expand our prompting evaluation beyond the zero-shot and one-shot setup from Micallef and Borg (2025), as we consider a larger number of shots in this work, up to ten.

2.3. Crowdsourcing in Sentiment Analysis

A large body of work has shown that crowdsourcing can deliver reliable sentiment analysis at scale. For tweet- and sentence-level polarity, a common pattern is to collect multiple independent judgement per item. Platforms like Appen or Amazon Mechanical Turk are commonly used, especially when the data is in a widely spoken language such as English. One such example is the SemEval-2017 Task 4 Sentiment Analysis in Twitter (Rosenthal et al., 2017) with the dataset annotated on CrowdFlower[3] and using a majority voting procedure.

Crowdsourcing has proved effective outside English and in subjective/figurative phenomena tightly coupled with sentiment. For Italian, Stranisci et al. (2016) use CrowdFlower to annotate sentiment in tweets of a political nature. They carry out an in-depth disagreement analysis, highlighting the need to model annotator subjectivity in contentious, politically polarised discourse. In Arabic, Abu Farha and Magdy (2020) re-annotated tweets for sarcasm, sentiment, and dialect via Figure-Eight (Appen), with language-based annotator gating and a minimum of 3 judgments per item. Within a domain-specific setting, Gabryszak and Thomas (2022) created the MobASA corpus, focusing on sentiment and social inclusion in the mobility domain. The process used combined expert labels with a crowdsourced subset (Crowdee) to scale coverage while maintaining quality on accessibility-related aspects.

While using existing third-party crowdsourcing platforms has its advantages due to the readily available infrastructure, an additional effort would be required to entice native speakers to register on such platforms to participate in the annotation exercise. In practice, due to the low-resource nature of the language, finding Maltese speakers on such platforms is challenging. This necessitates the development of a customised annotation platform (Section 3.3) to facilitate community-focused crowdsourcing efforts.

[3] CrowdFlower evolved into Figure8, and is now Appen.

3. The SentiMalti Dataset

This section outlines the construction of the new Sentiment Analysis dataset – **SentiMalti** – constructed from a combination of data from prior work and our newly annotated data. Section 3.1 describes the data sources that were scraped. We carried out a number of preprocessing steps, including anonymisation, and describe this in Section 3.2. Section 3.3 details the crowdsourcing campaign, including the design and development of our crowdsourcing annotation platform. After the crowdsourcing campaign, we conduct additional quality checks and filtering before finalising the dataset as explained in Section 3.4.

3.1. Data sources

In an effort to have a diverse representation of the Maltese culture, different data sources were selected.

YouTube Three different channels were identified in order to scrape comments and reactions by users. These are Djun MT,[4] Ricky Caruana Podcast,[5] and Jon Mallia Podcast.[6] The Jon Mallia and Ricky Caruana podcasts were selected for their interviews with a variety of guests from different backgrounds. This diversity naturally leads to a broader spectrum of discussions, generating positive, negative and neutral comments across multiple domains. The Djun channel was included to introduce more diversity into the dataset, as it features local music content that attracted both positive and critical comments from viewers. The comments were scraped systematically using the Youtube API.

Facebook A group called RUBS (Are you being served).[7] Comments were collected through a dedicated scraper built using Selenium WebDriver. This group was chosen for the variety of posts, including praise or criticism of products and services in Malta. All names were anonymised.

Reddit r/malta channel.[8] This platform was chosen as a secondary source because it hosts informal discussions on a wide range of topics relevant to the Maltese community. However, since a majority of discussions on r/Malta are conducted in English, a subset of posts was manually selected where Maltese was actively used. These posts typically included discussions about local events, politics, or cultural matters. Comments were extracted using the Reddit API.

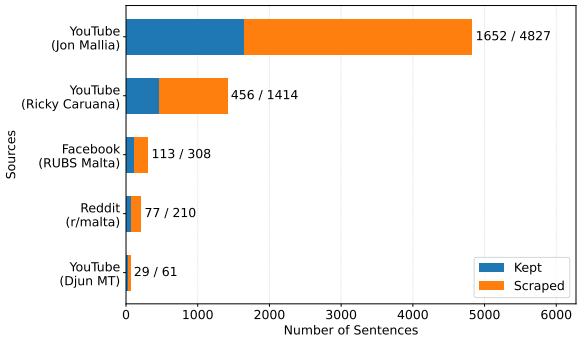


Figure 1: Data distribution of the newly collected data.

These three primary sources resulted in different quantities of comments. Initial scraping captured all comments, and a post-processing step then filtered out comments in English using a language identifier.

In total, 6,820 sentences were scraped. Figure 1 shows the data distribution from these different sources.

3.2. Data Preprocessing

All data collected from the scrapers was processed through a unified, sequential preprocessing pipeline. Each stage transforms the data and passes its output to the next, enabling step-by-step verification.

Text Cleaning This first stage normalises text from all sources. Operations include standardising brackets and quotes, converting common punctuation emojis to text (e.g., ⁉ to !?), normalising newlines and ellipses (e.g., "....." and ".." to "..."), limiting consecutive punctuation (e.g., "!?!?" to "!?", "????" to "??"), and correcting spacing around punctuation. Moreover, YouTube comments included timestamps that referenced specific moments in the videos (e.g., "01:23" or "2:15"). These timestamps were automatically removed during pre-processing using a custom regular expression. Cleaning occurs first to establish a normalised text base, which is necessary for accurate pattern-matching and consistent sentence boundary detection in the subsequent steps.

Text Anonymisation This stage replaces Personally Identifiable Information (PII) with generic placeholders. Specifically, it targets URLs, email addresses, phone numbers, and usernames (e.g., strings starting with "@"). Detection patterns are

[4] https://www.youtube.com/@DjunMT
[5] https://www.youtube.com/@rickycaruanapodcast/
[6] https://www.youtube.com/@JonMalliaPodcast
[7] https://www.facebook.com/groups/RUBS.Malta
[8] https://www.reddit.com/r/malta/

applied with defined precedence and length constraints to reduce false positives. We additionally match common names and surnames using a predefined list obtained from Gianola et al. (2020). Anonymisation is performed after cleaning to operate on normalised text and before sentence splitting, as PII often contains punctuation that could otherwise disrupt sentence boundaries. The use of placeholders also simplifies the splitting process.

Sentence Splitting This stage segments text into individual sentences. It employs a protect-split-restore process: first, it tokenises anonymisation placeholders, numbers with internal punctuation, ellipses, and specified Maltese/English abbreviations (e.g., "Dr.", "Mr.", "Prof."). Then, the system splits the protected text based on sentence-ending punctuation and newlines. Finally, it reverts temporary tokens. Accurate sentence splitting requires cleaned and tokenised text and is a prerequisite for sentence-level language filtering.

Language Filtering This step identifies and retains Maltese sentences. We use the langid library (Lui and Baldwin, 2012), configured for English/Maltese classification, classifying text as Maltese if its English probability is below an empirically determined 0.94 threshold. This approach was selected to account for code-switching, as online Maltese text frequently contains English words or phrases. Applying language filtering after sentence splitting ensures effective operation on clearly delineated sentences, free of formatting or PII that could skew identification.

Data Combination This final stage aggregates all processed Maltese sentences from all platforms into one consolidated file. Duplicate sentences are removed. The system saves each unique sentence with a sequential ID, its source identifier, and its text content.

3.3. Crowdsourcing Sentiment Annotations

In order to crowdsource the annotations, we built a custom website.[9] The website first displays a landing page, shown in Figure 2, which consists of an explanation of the exercise and provides a statement clarifying that no personal information is associated to the responses and that the participant can choose to stop at any point of the annotation exercise.[10]

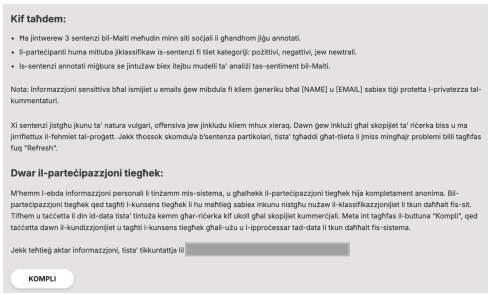


Figure 2: Website landing page

Figure 3: Website annotation page

The participant could then choose to proceed to the annotation page, shown in Figure 3. The participant is presented with 3 sentences to annotate using the provided colour-coded buttons underneath each sentence. The buttons correspond to our target labels: Positive, Negative, or Neutral. Since sentences are presented out of their original context, a participant might be unsure of the sentiment in certain instances. Hence, we also included a fourth button marked Unsure (*Ma nafx*). By doing so, we wanted to minimise random selection in cases where the participant might have felt 'pressured' to select a label, thus allowing some flexibility in the participant's judgement.

The website displays sentences in a random order, however it prioritises comments with fewer than three annotations and no majority sentiment. Within a session, the participant can annotate a comment only once.[11] Once a comment gets three annotations and a majority vote, it is no longer shown. If no majority is reached, it remains available to users who haven't yet annotated it.

Participants were recruited through our institutional mailing list and through social media posts. We specifically targeted Maltese language-related groups where people often discuss Maltese words, translations and other language-specific questions.

[9] https://nlpgroup.research.um.edu.mt/sentiment/

[10] Note that ethical approval was provided by the institution's ethics board.

[11] Note that since participants do not register, we can only track which sentences were annotated in a given session. If the same participant enters the website on another day, this is considered as a new participant/session.

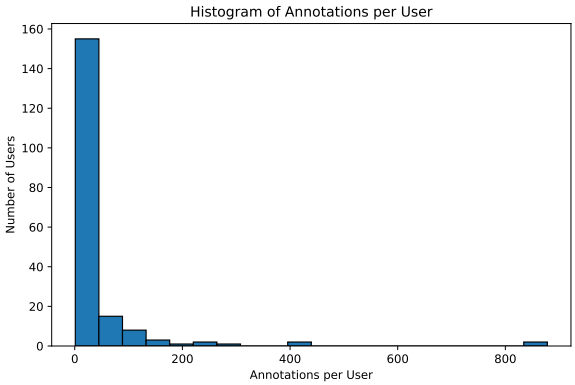


Figure 4: Distribution of the number of annotations per user.

After the crowdsourcing campaign, we received annotations from 878 sessions, with a total of 7,843 annotations. As shown in Figure 4, most participants annotated a few sentences, although 17 participants annotated upwards of 100 different sentences. The mean and median annotations per session were 41.50 and 15, respectively.

3.4. Final Dataset Construction

After the crowdsourcing campaign, we manually reviewed a subset of sentences for quality purposes. All sentences with at least one unsure vote were rechecked to confirm that the majority vote should stand. We also reviewed sentences where we had more than 3 votes but the agreement was less than 66% (e.g. sentences that had 5 votes: 3 voting one class, 1 voting a second class, and 1 voting a third class) and sentences with only 2 votes from the crowdsourcing exercise. In all these cases, if the first reviewer did not agree with the majority vote, the sentence would be discussed with a second reviewer, and a decision on the final label would be made. In total, we reviewed 268 sentences, of which 57 were reviewed by the second reviewer as well.

A total of 37 sentences were manually removed during the review process. These included sentences that were highly ambiguous in terms of sentiment polarity (e.g. *Dan il-podcast ġabli d-dmugħ.* 'This podcast brought tears [to my eyes].'), or they were ambiguous due to incorrect orthography and both reviewers were unsure of the meaning of the sentence.

The final dataset consists of 2,327 sentences, with sentiment distributions of 914 positive (39%), 725 negative (31%), and 688 neutral (30%). As shown in Figure 5, the newly collected data is significantly larger than the previous data sources, with some skewness towards positive rather than negative. We split the data into train, validation, and test using similar ratios as Martínez-García et al. (2021), that is, 70%, 10%, and 20%, respectively.

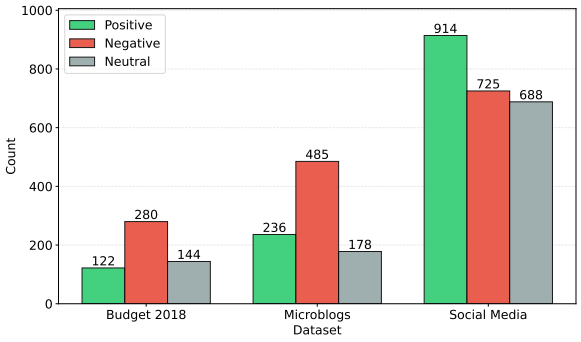


Figure 5: Label distribution for each subset.

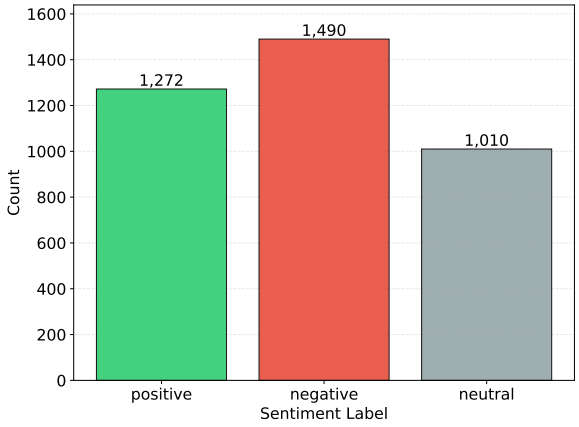


Figure 6: Label distribution of the entire data.

Subset	Train	Validation	Test
Budget 2018	389	50	107
Microblogs	621	94	184
Social Media	1,628	232	467
All	2,638	376	758

Table 1: Distribution of the combined data

In addition, we revisited the original datasets by Cortis and Davis (2019) and Dingli and Sant (2016) and combine them with our newly collected data, to construct **SentiMalti**. For sentences already included in the dataset from Martínez-García et al. (2021), we ensure that they remain in the same splits, splitting the remaining sentences randomly with the same proportion 70%-10%-20% as before. This allows us to include neutral sentences that were not included by Martínez-García et al. (2021).

Figure 6 shows the label distribution of the combined SentiMalti dataset. The combined dataset shows a more or less balanced label distribution, bridging the positive and negative classes closer to each other, thus improving upon the previously existing class imbalance. Table 1 provides the distribution of the datasets split into train, validation, and test.

4. Language Model Evaluation

We perform an evaluation of a variety of models on the newly constructed dataset. We decided to experiment with both encoder-only models and instruction-tuned open source models to better understand the difference in performance, if any, when these type of models are applied to a low-resource language.

4.1. Experimental setup

We consider two types of models, which we either fine-tune or prompt.

Fine-tuned For this setup, we consider encoder models whereby we perform parameter updates by fine-tuning them on the training data. In addition to training on **All** of the data, we consider two experimental variants where models are trained on the new **Social Media** data only or the previous data (**Budget 2018 + Microblogs**) only. Since all models considered are BERT-based, we add a classification head on the language model to output one of the three labels, and training is performed using the Transformers library (Wolf et al., 2020). Models are trained for a maximum of 200 epochs using early stopping on the corresponding validation set with a patience of 20 epochs. We use a batch size of 16, a classifier dropout of 0.1, a learning rate of 2e-5 with an inverse square-root learning rate scheduler and an AdamW optimiser, a warmup of 1 epoch, and a weight decay of 0.05. The models that we fine-tune are the following:

- **BERTu** (Micallef et al., 2022): A BERT-based model pre-trained from scratch on Maltese data with 126 million parameters.
- **Glot500** (Imani et al., 2023): A model based on XLM-R (Conneau et al., 2020) further pre-trained on 511 languages, including Maltese, having a total of 395 million parameters.

In Appendix A, we also present results with other models, including mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mBERTu (Micallef et al., 2022), and mmBERT (Marone et al., 2025). BERTu and Glot500 generally perform better than these models, except for mBERTu, which benefits from its significant Maltese pre-training.

Prompted These models are prompted to produce classification outputs without performing any parameter updates. Models are given the sentence together with an instruction, formatted using a prompt template. In addition to the zero-shot setup, we consider a few-shot setup, where we prepend examples sampled from the appropriate training set to aid the model learn in context from the provided demonstrations. We follow the experimental setup from Micallef and Borg (2025), including their prompt template and adding an additional neutral class. To get the target label, we compute the log-likelihood of each label given the prompt and choose the one with the highest log-likelihood. All prompting experiments are conducted using the LM Evaluation Harness library (Gao et al., 2024). The models that we prompt are the following:

- **Aya-101** (Üstün et al., 2024): A 13 billion parameter encoder-decoder model with multilingual pre-training and instruction-tuning, including a proportion of Maltese data.
- **Gemma 2 Instruct 9B** (Gemma Team et al., 2024): A decoder-only model with 9 billion parameters. Its training data is multilingual, although there are no details on whether Maltese is included.

Macro-averaged F1 is used as the primary evaluation metric. Each experiment is done 5 times with different random seeds, and we report the average and standard deviation.

4.2. Results

The results are shown in Table 3 where we report the scores on **All** of the data as well as the individual data subsets: **Budget 2018**, **Microblogs**, and **Social Media**.

Fine-tuned models BERTu, when fine-tuned on all the data, achieves the best score across all fine-tuning setups. This is largely due to the model's pre-training focus on Maltese, as BERTu performs better than Glot500 on all training setups and test sets, except for the Budget 2018 test set when trained on the previous data only and all the data. When training only on the new data, we observe better performance than when training only on the previous data. This is largely due to the better performance on the new data, which makes up a larger proportion of the overall test data. In contrast, when training only on previous data, better performance is observed for the Budget 2018 and Microblogs subsets, since the domain and style of the data are the same between training and testing. Despite this, when combining all training datasets together, better performance is achieved across test sets, since the model is exposed to a larger variety of data and is able to generalise better.

Prompted models Both models perform on par, with the largest performance gap noticeable in one-shot. Gemma 2 Instruct 9B performs consistently better than Aya-101 on the Budget 2018 subset, while performance is more or less on par on the

Model	Budget 2018	Microblogs	Social Media	All (SentiMalti)
Fine-Tuning with Previous Data Only (Budget 2018 + Microblogs)				
BERTu	56.93 ± 2.73	**65.35 ± 2.75**	60.63 ± 1.25	62.17 ± 0.32
Glot500	61.33 ± 2.61	56.83 ± 1.46	51.82 ± 1.10	54.92 ± 1.27
Fine-Tuning with New Data Only (Social Media)				
BERTu	49.10 ± 3.05	**65.55 ± 1.94**	**68.25 ± 0.99**	66.06 ± 1.00
Glot500	42.57 ± 4.32	52.65 ± 1.84	59.44 ± 1.93	56.55 ± 1.63
Fine-Tuning with All (SentiMalti) Data (Budget 2018 + Microblogs + Social Media)				
BERTu	57.65 ± 1.43	**67.71 ± 2.64**	**68.88 ± 0.37**	**68.36 ± 0.62**
Glot500	59.42 ± 3.30	58.56 ± 2.30	61.67 ± 1.42	62.00 ± 1.52
Zero-Shot Prompting				
Aya-101	56.38 ± 0.00	61.17 ± 0.00	58.87 ± 0.00	60.27 ± 0.00
Gemma 2 Instruct 9B	60.94 ± 0.00	**63.42 ± 0.00**	56.82 ± 0.00	60.50 ± 0.00
One-Shot Prompting				
Aya-101	58.10 ± 0.96	**62.51 ± 2.47**	60.17 ± 0.64	61.15 ± 2.04
Gemma 2 Instruct 9B	**61.97 ± 6.09**	60.60 ± 1.51	62.64 ± 0.77	63.94 ± 0.80
Five-Shot Prompting				
Aya-101	64.14 ± 1.83	**66.25 ± 2.60**	**68.35 ± 1.58**	**68.65 ± 0.82**
Gemma 2 Instruct 9B	**69.69 ± 1.90**	**66.91 ± 1.27**	66.51 ± 0.77	**68.15 ± 1.05**

Table 2: Evaluation Results in terms of macro-averaged F1 (based on 5 runs with different random seeds). The best average score for each test set is shown in **bold and underline**. We also perform a one-tailed t-test against the best model, and show the results in **bold** for scores that are not found to be significantly worse (with a p-value = 0.05 and Bonferroni correction).

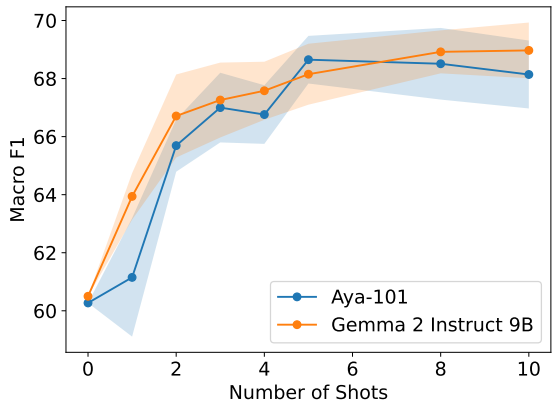


Figure 7: Performance on all the data with increasing number of shots.

other subsets. Both models achieve the best overall score with five-shot. In addition to these results, we experimented with a larger variety of shots, up to 10, and visualised these results in Figure 7. Between zero-shot and two-shot, a large performance improvement is observed, and similar to Zhang et al. (2024), diminishing returns are observed thereafter. Performance plateaus at five-shot, as with larger number of shots Gemma 2 Instruct 9B improves slightly while Aya-101 does slightly worse.

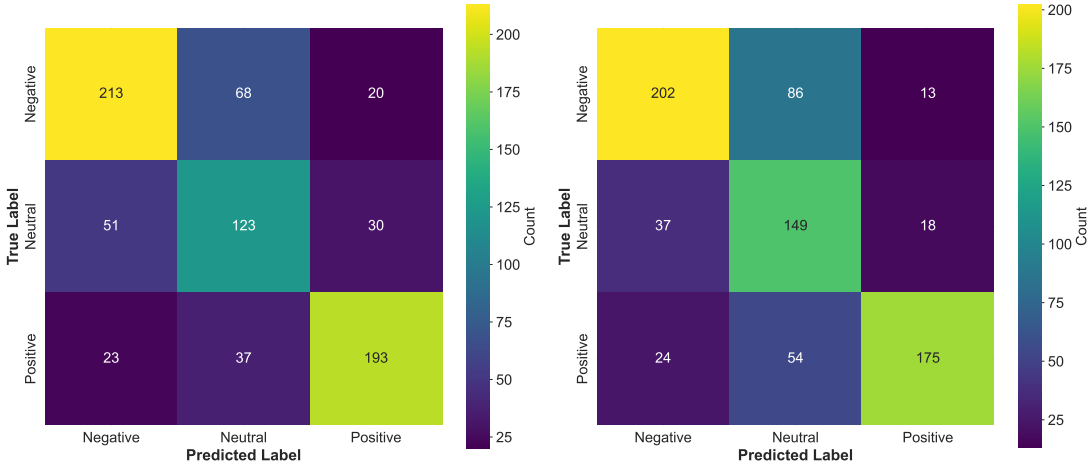
Overall, the best performance is obtained with both prompted models in five-shot and BERTu when fine-tuned on all of the data. Figure 8 provides a visual representation of the predictions for the best performing models. BERTu is able to predict positive and negative classes more accurately, while Aya-101 is better at classifying neutral instances. However, Aya-101 over-predicts instances as neutral, resulting in a high recall for the neutral class but lower precision for the other classes. BERTu also misclassified positive and negative instances as neutral, but to a lesser extent, as it also misclassified a portion of neutral instances as positive and negative.

5. Conclusion

We introduce SentiMalti, a Maltese social media sentiment resource and a reproducible pipeline from scraping to Maltese-aware preprocessing and crowdsourced annotation. The final dataset comprises 2,327 three-way labelled sentences and, when merged with prior Maltese corpora, yields a 3,772-instance benchmark with standard splits for comparative evaluation.

We also included an evaluation of fine-tuned and prompted language models. Our experiments show that few-shot prompting with Aya-101 matches, and slightly exceeds, the strongest fine-tuned encoder (BERTu) on the full test set (68.65 vs. 68.36 macro-F1). Confusion-matrix analysis indicates that BERTu is more precise on polarised classes, while Aya-101 achieves higher recall on neutral, suggesting complementary error profiles that future ensembles or calibration could exploit.

Practically, five-shot prompting offers a strong,



(a) BERTu Fine-Tuned with All Data

(b) Aya-101 Five-Shot Prompting

Figure 8: Confusion Matrix on the combined test data for a selection of models.

data-efficient baseline, while a fine-tuned BERTu remains competitive and less computationally expensive for inference. The resources, code, and released model are intended to strengthen the resources available for Maltese NLP, enabling further research on code-switching robustness, strategies for annotation, and domain transfer.

6. Ethics Statement

Ethics approval was obtained from the institution's ethics board. We scraped publicly available comments, except for 113 comments from the Facebook group RUBS. No information was retained regarding comment usernames or timestamps. Moreover, comments were anonymised, and any names were replaced with a `<PERSON>` tag. Participants in the annotation process were recruited on a voluntary basis through an institution's newsletter and a social media post. We ensured that no identifiable information about participants was retained. The website and recruitment notices were all published in Maltese. No English translation was available since this exercise targeted Maltese native speakers.

7. Bibliographical References

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

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. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. 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.

Alexei Dingli and Nicole Sant. 2016. Sentiment analysis on Maltese using machine learning. In *The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016)*, pages 21–25.

Aleksandra Gabryszak and Philippe Thomas. 2022. MobASA: Corpus for aspect-based sentiment analysis and social inclusion in the mobility domain. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*,

pages 35–39, Marseille, France. European Language Resources Association.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, et al. 2024. The language model evaluation harness.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size.

Lucie Gianola, Ēriks Ajausks, Victoria Arranz, Chomicha Bendahman, Laurent Bié, Claudia Borg, Aleix Cerdà, Khalid Choukri, Montse Cuadros, Ona de Gibert, et al. 2020. Automatic Removal of Identifying Information in Official EU Languages for Public Administrations: The MAPA Project. In *Legal Knowledge and Information Systems*, volume 334 of *Legal Knowledge and Information Systems*, pages 223–226, Brno, Prague, Czech Republic. IOS Press.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Kurt Micallef and Claudia Borg. 2025. MELABenchv1: Benchmarking large language models against smaller fine-tuned models for low-resource Maltese NLP. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20505–20527, Vienna, Austria. Association for Computational Linguistics.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pretraining data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for 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, Vancouver, Canada. Association for Computational Linguistics.

Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2892–2899, Portorož, Slovenia. European Language Resources Association (ELRA).

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7342–7371, Bangkok, Thailand. Association for Computational Linguistics.

Model	Budget 2018	Microblogs	Social Media	All (SentiMalti)
Fine-Tuning with Previous Data Only (Budget 2018 + Microblogs)				
BERTu	56.93 ± 2.73	65.35 ± 2.75	60.63 ± 1.25	62.17 ± 0.32
mBERTu	61.07 ± 2.62	58.63 ± 1.22	54.66 ± 1.83	57.12 ± 1.31
mBERT	48.00 ± 3.31	45.07 ± 2.25	39.84 ± 2.63	42.97 ± 2.30
XLM-R	52.33 ± 1.87	43.52 ± 3.16	42.24 ± 1.69	44.79 ± 1.58
Glot500	**61.33 ± 2.61**	56.83 ± 1.46	51.82 ± 1.10	54.92 ± 1.27
mmBERT	55.81 ± 3.15	56.91 ± 4.57	47.31 ± 0.39	51.23 ± 1.26
Fine-Tuning with New Data Only (Social Media)				
BERTu	49.10 ± 3.05	65.55 ± 1.94	68.25 ± 0.99	66.06 ± 1.00
mBERTu	49.99 ± 2.74	60.23 ± 1.92	66.15 ± 0.97	63.59 ± 1.46
mBERT	37.42 ± 3.06	49.45 ± 1.81	52.63 ± 2.06	50.88 ± 1.58
XLM-R	38.58 ± 5.60	43.03 ± 1.62	55.33 ± 1.06	50.67 ± 0.89
Glot500	42.57 ± 4.32	52.65 ± 1.84	59.44 ± 1.93	56.55 ± 1.63
mmBERT	52.01 ± 2.55	56.16 ± 2.23	59.04 ± 1.47	58.71 ± 0.78
Fine-Tuning with All (SentiMalti) Data (Budget 2018 + Microblogs + Social Media)				
BERTu	57.65 ± 1.43	**67.71 ± 2.64**	**68.88 ± 0.37**	**68.36 ± 0.62**
mBERTu	55.40 ± 2.81	62.64 ± 0.69	67.31 ± 0.70	65.83 ± 0.27
mBERT	49.27 ± 1.07	50.23 ± 1.27	53.02 ± 1.56	53.48 ± 0.92
XLM-R	58.57 ± 3.65	47.39 ± 3.20	57.53 ± 1.77	57.43 ± 1.34
Glot500	59.42 ± 3.30	58.56 ± 2.30	61.67 ± 1.42	62.00 ± 1.52
mmBERT	60.73 ± 3.11	58.13 ± 1.86	60.14 ± 1.22	61.24 ± 1.18

Table 3: Results of all fine-tuned encoder-only (including the ones from Table 3) in terms of macro-averaged F1 (based on 5 runs with different random seeds). The best average score for each test set is shown in **bold and underline**.

A. Results on additional encoder-only models

In addition to BERTu (Micallef et al., 2022) and Glot500 (Imani et al., 2023), we include results for mBERTu (Micallef et al., 2022), multilingual BERT (mBERT) (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and modern multilingual BERT (mmBERT) (Marone et al., 2025). The results of all fine-tuned models, including the ones presented in Section 4, is shown in Table 3.