Proceedings of SIGUL2022 @LREC2022, pages 161–168
Marseille, 24-25 June 2022

Baseline English and Maltese-English Classification Models for Subjectivity Detection, Sentiment Analysis, Emotion Analysis, Sarcasm Detection, and Irony Detection

Keith Cortis, Brian Davis
ADAPT Centre
Dublin City University
Glasnevin, Dublin 9, Ireland
{keith.cortis, brian.davis}@adaptcentrie.ie

Abstract

This paper presents baseline classification models for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection. All models are trained 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. Traditional supervised algorithms namely, Support Vector Machines, Naïve Bayes, Logistic Regression, Decision Trees, and Random Forest, are used to build a baseline for each classification task, namely subjectivity, sentiment polarity, emotion, sarcasm, and irony. Baseline models are established at a monolingual (English) level and at a code-switched level (Maltese-English). Results obtained from all the classification models are presented.

Keywords: opinion mining, social media, subjectivity analysis, sentiment analysis, emotion analysis, irony detection, sarcasm detection, social data, code-switching

1. Introduction

Finding out what other people think about a product or service has always been a very important part of an individual's and/or organisation's information gathering behaviour especially during a decision making process. Before the World Wide Web awareness, people asked their friends and colleagues about recommendations for an automobile mechanic, or about whom they plan to vote for in the upcoming elections, and checked with the consumer reports before buying a house appliance. Organisations usually conducted market analysis in the form of opinion polls, surveys, and focus groups in order to capture public opinion concerning their products and services (Liu, 2010). The advent of the Social Web and the massive increase of user-generated content posted on social media platforms and newswires commenting sections, allows users to create and share content and their opinions directly to the public, thus circumventing possible forms of bias (by acquaintance of experts only). Such user-generated content is invaluable for certain needs, such as improving an entity's service or perception and tracking citizen opinion to aid policy makers and decision takers (Hilts and Yu, 2010). Opinion-rich resources have been growing both in terms of availability and popularity.

The year of 2001 marked the beginning of widespread awareness of the research problems and opportunities for Opinion Mining and Sentiment Analysis (Pang and Lee, 2008). Online review sites and personal blogs were early examples of such opinionated resources, whereas social networking (e.g., Facebook[1]), microblogging (e.g., Twitter[2]), travel (e.g., TripAdvisor[3]), and newswire (e.g., Reuters[4]) services are nowadays the most popular. This created new opportunities and challenges for Opinion Mining, especially on user-generated content spread across heterogeneous sources, such as newswires and social networking services.

This paper presents baseline classification models for **five** opinion classification tasks: *subjectivity detection*, *sentiment analysis*, *emotion analysis*, *sarcasm detection*, and *irony detection*. These are based on 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.

In terms of language, this social data is in one of the following languages: English —a high-resourced language, Maltese —a low-resourced language, and Maltese-English —a code-switched language. Baseline models are established at a monolingual level using user-generated content in English, and at a code-switched level using user-generated content in Maltese-English and Maltese. Section 2 presents a review of social datasets available for Opinion Mining, the algorithms generally used for evaluating them, and other relevant studies within this research area. The experiments carried out to establish the baseline models are discussed in Section 3, with some conclusions and future work presented in Section 4.

[1] `https://www.facebook.com`

[2] `https://www.twitter.com`

[3] `http://www.tripadvisor.com`

[4] `https://www.reuters.com`

2. Related Work

Studies focusing on text classification tasks, such as sentiment analysis, at a binary (two classes) and/or multi-class (more than two classes) level generally use machine learning (ML) and deep learning (DL) supervised algorithms for building their baseline models. A Social Opinion Mining systematic review (Cortis and Davis, 2021b) analysed a large number of studies that make use of social data, such as user-generated content from social media platforms, and identified techniques used for carrying out classification tasks in this research area. In terms of traditional supervised learning algorithms, the most common ones used for baseline, experimentation, evaluation and/or comparison purposes are Naïve Bayes (NB) (Lewis, 1998), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Logistic Regression (LR) (McCullagh, 1984) / Maximum Entropy (MaxEnt)–generalisation of LR for multi-class scenarios (Yu et al., 2011), Decision Tree (DT) (Quinlan, 1986), and Random Forest (RF) (Breiman, 2001). The choice of traditional supervised learning algorithms selected is supported by other Opinion Mining reviews, such as (Ravi and Ravi, 2015), (Hemmatian and Sohrabi, 2019), (Carvalho and Plastino, 2021), (Ligthart et al., 2021). Even though recent advances in Opinion Mining has seen an increase in the use of DL approaches, such as the Transformer model architecture (Vaswani et al., 2017), traditional ML algorithms are still very much used to carry out Opinion Mining classification tasks, with good results obtained especially on small datasets (Ligthart et al., 2021).

Several high-quality Opinion Mining social datasets are available for research purposes as part of shared evaluation tasks, such as the International Workshop on Semantic Evaluation (SemEval)[5] and/or through open access repositories, such as Zenodo[6]. Teams submitting their systems in the SemEval sentiment analysis task on code-mixed tweets (Patwa et al., 2020) used the following techniques, traditional ML algorithms such as NB, LR, RF, and SVM; word embeddings such as word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Joulin et al., 2016); and DL algorithms such as Convolutional Neural Network (CNN) (LeCun et al., 1990), and Bidirectional Enconder Representations from Transformers (BERT) (Devlin et al., 2018). In (Gupta et al., 2017), several ML (SVM best performer) and DL algorithms are used as baselines for contextual emotion detection on tweets.

A baseline SVM system was trained in numerous SemEval tasks, such as (Mohammad et al., 2018) for affect in tweets and (Pontiki et al., 2016) for aspect-based sentiment analysis. Similarly, SVM performed well on irony detection (Van Hee et al., 2018) and sentiment analysis (Rosenthal et al., 2017) in tweets. Participants in the SemEval task focusing on fine-grained sentiment analysis on financial microblogs and news (Cortis et al., 2017) made use of lexicon-based, ML, DL, and hybrid techniques, similar to (Patwa et al., 2020). An approach based on SVM was used in (Kothari et al., 2013) for subjectivity classification of news articles' comments and tweets. In (Appidi et al., 2020), ML algorithms such as SVM were used for emotion classification experiments on an annotated corpus of code-switched Kannada-English tweets. Bansal et al. used SVM and RF for training baseline models to show how code-switching patterns can be used to improve several downstream Natural Language Processing (NLP) applications (Bansal et al., 2020). In (Mamta et al., 2020), the authors also implemented baseline models for sentiment analysis using ML and DL algorithms, such as SVM and CNN. Similarly, the authors in (Yimam et al., 2020) built several baseline models for Amharic sentiment analysis from social media text using ML algorithms, such as SVM and LR.

3. Experiments

In this paper, we experimented with multiple classification models catering for the English, Maltese, and Maltese-English languages across **five** different social opinion dimensions, namely *subjectivity*, *sentiment polarity*, *emotion*, *irony*, and *sarcasm*. All experiments have been carried out in the Python using Jupyter Notebook[7] on a machine with an Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz processor and 8.00 GB (7.88 GB usable) installed memory (RAM).

3.1. Baseline Models

Baseline models for each social opinion dimension were built using the following eight supervised learning algorithms:

- **NB**: Multivariate Bernoulli NB (MBNB)–classifier suitable for discrete data and is designed for binary/boolean features (scikit learn, a), and Complement NB (CNB)–designed to correct "severe assumptions" made by the standard Multinomial NB classifier and suited for imbalanced datasets (scikit learn, b);
- **SVM**: Support Vector Classification (SVC)–C-SVC implementation based on libsvm (a library for SVM) (scikit learn, h), Nu-Support SVC (NuSVC)–similar to SVC however can control the number of support vectors (scikit learn, f), and Linear SVC–similar to SVC however has more flexibility and supports both dense and sparse input (scikit learn, d);
- **LR**: a probabilistic classifier also known as logit or Maximum Entropy (scikit learn, e);
- **DT**: an optimised version of the Classification and Regression Trees (CART) algorithm (scikit learn, c); and
- **RF**: an ensemble of decision tree algorithms (scikit learn, g).

[5] https://semeval.github.io/

[6] https://zenodo.org/

[7] https://jupyter.org/

The scikit-learn[8] ML library was used for building the baseline models. This consists of a set of tools for data mining and analysis, such as pre-processing, model selection, classification, regression, clustering, and dimensionality reduction.

3.2. Approach

The Opinion Mining approach for building baseline models consists of the following steps, namely data acquisition, pre-processing, model generation, and model evaluation.

3.2.1. Dataset

The dataset of multidimensional and multilingual social opinions for Malta's Annual Government Budget[9] (Cortis and Davis, 2021a) has been 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 three newswires, namely Times of Malta[10], MaltaToday[11], and The Malta Independent[12], and one social networking service, namely Twitter. In terms of languages, the majority of the online posts were in English (74.09%) with most of the rest being in Maltese-English or Maltese (24.99%). It is important to note that the online posts in Maltese-English and Maltese have been merged together due to the low amount of online posts in Maltese only. 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-English and Maltese (columns 6 and 7).

3.2.2. Pre-processing

Pre-processing on the online posts used for building the baseline models was carried out, using the following NLP tasks of a syntactic nature:

- **Data cleaning**: Removal of any numbers, HTML/XML tags, special characters and whitespaces;
- **Tokenisation**: text composed of string of words or sentences split into tokens, in terms of alphabetic and non-alphabetic characters, using the NLTK (Bird et al., 2009) word punctuation tokeniser;
- **Stemming**: removes suffices or prefixes used with a word to reduce inflectional forms to a common base form, using NLTK's implementation of the Porter stemming algorithm[13]; and
- **Conversion of textual data into numerical representations**: term frequency and inverse document frequency (TF-IDF) (Salton and McGill, 1986) statistical measure (using the scikit-learn TfidfVectorizer function) used to evaluate the word relevance in online posts and hence represent the online posts into a feature vector for training a classifier using any algorithm discussed in Section 3.1.

3.2.3. Model Generation

Given that the dataset used is relatively small in terms of data volume, we are not in a position to omit a chunk of data for model generation. Therefore, cross-validation provides us with a better modelling approach for small datasets, as opposed to the traditional training-validation-test set split. Stratified 10-fold cross-validation is applied on the entire dataset being used for model generation and evaluation. This cross-validation technique is used since the ratio between the target classes is preserved as is in the full dataset. It is also adequate for imbalanced datasets such as the one being used, as reflected in Table 1. Moreover, this technique just shuffles and splits the dataset once into 10 folds. Therefore, the test sets used for validating the trained model (on k - 1 of the folds used as training data) do not overlap between any of the 10 splits. Lastly, the model itself is trained 10 times, with the weights and any biases being reset with each new model. This cross-validation procedure was applied for each baseline model built using the supervised learning algorithms discussed in Section 3.1. Baseline classification models for *subjectivity*, *sentiment polarity*, *emotion*, *sarcasm*, and *irony*, were built on i) the subset of English online posts and ii) the subset of Maltese-English and Maltese online posts.

3.3. Results and Discussion

Results of the baseline classification models mentioned in Section 3.2.3 are presented and discussed in this section. Table 2 displays results obtained on the subset of English online posts, whereas Table 3 displays results obtained on the subset of Maltese-English and Maltese online posts (merged together due to the low amount of online posts in Maltese only).

The following evaluation metrics were used to measure the classification performance of the models generated for each social opinion dimension:

- **F1 score weighted** (Chinchor, 1992): 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,

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

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

[10] https://www.timesofmalta.com/

[11] https://www.maltatoday.com.mt/

[12] https://www.independent.com.mt/

[13] https://tartarus.org/martin/PorterStemmer/

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

Table 1: Class distribution for each annotation per dataset

that is, the number of true instances for each label. This metric caters for label imbalance.

- **Balanced accuracy** (Brodersen et al., 2010): defined as the average of recall scores obtained per class. This metric is used for imbalanced binary and multi-class classification.

Both tables present the mean and standard deviation F1 score (weighted) and balanced accuracy results obtained for all eight supervised learning algorithms using the stratified 10-fold cross-validation technique.

With respect to the English data, the LR algorithm obtained the best F1 score (weighted) results for the subjectivity and irony classification models. The SVC and RF obtained the same results for the latter model. The CNB algorithm produced the best F1 score (weighted) for the sentiment polarity and emotion classification models, whereas NuSVC fared best for the sarcasm classifier. When considering the balanced accuracy, the CNB algorithm produced the best results for all the social opinion dimensions.

As for the results on the Maltese-English and Maltese data, the CNB algorithm fared best in terms of F1 score (weighted) for the subjectivity, emotion (same as for English data), and irony classification models. The LinearSVC algorithm produced the best F1 score (weighted) for the sentiment polarity classifier, whereas the LR, SVC, and RF algorithms obtained the best and same results for sarcasm. Similar to the results obtained on the English data, the CNB algorithm produced the best balanced accuracy results for subjectivity, sarcasm, and irony. On the other hand, LinearSVC obtained the best balanced accuracy results for sentiment polarity, whereas RF fared best for emotion.

The following are some observations on the results obtained:

- The CNB algorithm obtained good performance for all languages and handled the imbalanced classes better than the other algorithms.

- Results obtained for the subjectivity and sentiment polarity classifiers are very promising for the English subset and Maltese-English and Maltese subset, even though the latter subset only amounts to 1596 online posts and the classes are not evenly balanced (for both subsets).

- Further evaluation using online posts unseen by the trained models is needed on the emotion, sarcasm, and irony classifiers to ensure that they are not biased towards the majority classes (Paduraru and Breaban, 2019), due to small amount of online posts available for the minority classes. Resampling techniques (Cateni et al., 2014; More, 2016) such as over-sampling and under-sampling can be used for handling such imbalances.

Opinion Dimension	LR	LinearSVC	NuSVC	SVC	BNB	CNB	DT	RF
Subjectivity								
F1 score (weighted)								
Mean	**0.883841**	0.879541	0.876273	0.496998	0.840135	0.883805	0.836563	0.8721
Standard Deviation	0.090688	0.076603	0.099753	0.000954	0.100051	0.077748	0.080797	0.09332
Execution time (sec)	0.366402	0.152594	133.560791	138.974958	0.056846	0.049864	3.255599	38.103134
Balanced accuracy								
Mean	0.866635	0.86671	0.855156	0.5	0.811524	**0.873531**	0.827453	0.85808
Standard Deviation	0.109981	0.095431	0.118991	0	0.11106	0.09623	0.099956	0.108103
Execution time (sec)	0.325131	0.147605	128.942503	138.723404	0.051897	0.038896	3.552703	32.025443
Sentiment Polarity								
F1 score (weighted)								
Mean	0.773488	0.766855	0.777319	0.390174	0.776828	**0.783019**	0.722608	0.763451
Standard Deviation	0.070612	0.054157	0.053882	0.000444	0.053359	0.073829	0.049837	0.065426
Execution time (sec)	4.067413	0.409448	173.671773	146.321687	0.063825	0.044364	4.606840	35.520771
Balanced accuracy								
Mean	0.722771	0.717044	0.739063	0.333333	0.727495	**0.766624**	0.685624	0.714724
Standard Deviation	0.077534	0.058836	0.059551	0	0.056173	0.075009	0.063942	0.070536
Execution time (sec)	3.950933	0.397628	173.383033	144.026906	0.041853	0.037899	4.462762	41.809123
Emotion								
F1 score (weighted)								
Mean	0.558523	0.573032	0.565908	0.246018	0.559174	**0.597985**	0.53082	0.538238
Standard Deviation	0.028066	0.04086	0.032799	0.000952	0.050814	0.059299	0.047546	0.048382
Execution time (sec)	12.094085	1.239142	249.117511	153.079795	0.119210	0.055034	5.528538	42.201759
Balanced accuracy								
Mean	0.247898	0.282854	0.268255	0.125	0.248973	**0.319283**	0.265121	0.237785
Standard Deviation	0.023119	0.025369	0.025894	0	0.034061	0.035876	0.032245	0.028137
Execution time (sec)	11.447332	0.969887	237.525686	134.157565	0.097378	0.045877	5.374176	42.332807
Sarcasm								
F1 score (weighted)								
Mean	0.9681	0.967914	**0.968939**	0.9681	0.956555	0.954114	0.961048	0.9681
Standard Deviation	0.000925	0.001536	0.001145	0.000925	0.023294	0.050832	0.007896	0.000925
Execution time (sec)	0.190490	0.132643	58.925297	8.298284	0.066821	0.050832	2.718095	19.604499
Balanced accuracy								
Mean	0.5	0.508466	0.509438	0.5	0.558462	**0.566324**	0.555348	0.5
Standard Deviation	0	0.018602	0.018891	0	0.055718	0.072913	0.042625	0
Execution time (sec)	0.186504	0.116689	63.924314	8.433631	0.054854	0.040890	2.684616	20.277674
Irony								
F1 score (weighted)								
Mean	**0.940496**	0.940348	0.940073	**0.940496**	0.917422	0.92766	0.934038	**0.940496**
Standard Deviation	0.000932	0.003619	0.000979	0.000932	0.046925	0.023662	0.018002	0.000932
Execution time (sec)	0.208476	0.151596	87.929416	17.331174	0.080325	0.050864	2.785665	27.825442
Balanced accuracy								
Mean	0.5	0.510847	0.49956	0.5	0.535921	**0.561683**	0.558459	0.502632
Standard Deviation	0	0.019195	0.00073	0	0.073153	0.038838	0.051882	0.007895
Execution time (sec)	0.209446	0.134637	88.003576	16.294276	0.053856	0.041888	2.464509	30.312636

Table 2: Classification model results - English dataset

4. Conclusions and Future Work

The paper discusses preliminary results of baseline classification models for subjectivity detection, sentiment analysis, emotion analysis, sarcasm detection, and irony detection. In this respect, language specific models for English (monolingual) and Maltese-English (code-switched Maltese-English and monolingual Maltese) have been built. Deep neural network language models like BERT shall be fine-tuned to adapt to new domains, transfer knowledge from one language to another, and build new classification models. In this regard, multiple neural-based classification models for subjectivity, sentiment polarity, emotion, sarcasm, and irony, at a multilingual level using user-generated content in English, Maltese, and Maltese-English have already been published in (Cortis et al., 2021). Models capable of understanding English and Maltese data, both being Malta's official languages, can be used by governments for policy formulation, policy making, decision making, and decision taking. Multidimensional Social Opinion Mining provides a nuanced voice to the citizens and residents of Malta and hence leaves a positive impact on society at large.

Opinion Dimension	**LR**	**LinearSVC**	**NuSVC**	**SVC**	**BNB**	**CNB**	**DT**	**RF**
Subjectivity								
F1 score (weighted)								
Mean	0.839627	0.841091	0.845513	0.371596	0.772777	**0.854936**	0.817842	0.842926
Standard Deviation	0.103584	0.092145	0.096013	0.002719	0.15322	0.105658	0.112311	0.141601
Execution time (sec)	0.140372	0.126745	19.511326	18.196656	0.075907	0.024654	0.911396	11.830028
Balanced accuracy								
Mean	0.843608	0.842783	0.847955	0.5	0.802879	**0.864388**	0.83255	0.847062
Standard Deviation	0.088045	0.08498	0.084147	0	0.114555	0.09073	0.105097	0.119309
Execution time (sec)	0.088763	0.091463	19.070196	18.066252	0.055128	0.040290	0.844383	11.074797
Sentiment Polarity								
F1 score (weighted)								
Mean	0.689206	**0.739622**	0.725397	0.433592	0.593306	0.724719	0.711952	0.720501
Standard Deviation	0.081683	0.096397	0.106766	0.001532	0.060966	0.102363	0.089007	0.10021
Execution time (sec)	3.230092	0.266272	23.310936	16.025399	0.043515	0.036038	1.418103	12.771962
Balanced accuracy								
Mean	0.562462	**0.638019**	0.618941	0.333333	0.449516	0.612975	0.601	0.619531
Standard Deviation	0.063686	0.091922	0.101573	0	0.049415	0.08929	0.087626	0.101987
Execution time (sec)	2.763433	0.185026	22.411216	16.399088	0.030229	0.027661	1.324603	15.214964
Emotion								
F1 score (weighted)								
Mean	0.376882	0.427224	0.375519	0.234498	0.314026	**0.432851**	0.403896	0.418661
Standard Deviation	0.034389	0.054136	0.049605	0.002648	0.035831	0.047303	0.070764	0.06
Execution time (sec)	8.288645	0.411707	32.969826	21.907470	0.056324	0.036504	1.851777	17.749398
Balanced accuracy								
Mean	0.205188	0.275239	0.241458	0.125	0.162991	0.254581	0.246939	**0.276573**
Standard Deviation	0.022074	0.062415	0.05161	0	0.02325	0.031079	0.061356	0.054324
Execution time (sec)	7.956497	0.351134	32.464855	22.532657	0.044598	0.040448	2.101381	19.586962
Sarcasm								
F1 score (weighted)								
Mean	**0.931012**	0.929747	0.930699	**0.931012**	0.921376	0.9097	0.915169	**0.931012**
Standard Deviation	0.004388	0.006875	0.004848	0.004388	0.016343	0.036802	0.023135	0.004388
Execution time (sec)	0.107805	0.094377	14.096916	2.055357	0.058842	0.040891	1.042049	8.552041
Balanced accuracy								
Mean	0.5	0.498684	0.499671	0.5	0.507068	**0.530167**	0.491371	0.499671
Standard Deviation	0	0.003947	0.000987	0	0.035031	0.087631	0.027111	0.000987
Execution time (sec)	0.090132	0.085380	15.818028	1.977244	0.043882	0.035901	0.930696	8.509591
Irony								
F1 score (weighted)								
Mean	0.874091	0.87929	0.872834	0.874091	0.855613	**0.884021**	0.880597	0.873464
Standard Deviation	0.00409	0.009558	0.005784	0.00409	0.031979	0.042444	0.031066	0.004947
Execution time (sec)	0.092754	0.089761	12.690590	3.377020	0.040492	0.044879	1.090016	10.739472
Balanced accuracy								
Mean	0.5	0.518964	0.49863	0.5	0.507704	**0.611068**	0.584066	0.506458
Standard Deviation	0	0.030468	0.003139	0	0.034832	0.057851	0.049402	0.013006
Execution time (sec)	0.076793	0.091754	14.075315	3.323173	0.038205	0.032164	1.082346	10.587719

Table 3: Classification model results - Maltese-English and Maltese dataset

5. Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

6. Bibliographical References

Appidi, A. R., Srirangam, V. K., Suhas, D., and Shrivastava, M. (2020). Creation of corpus and analysis in code-mixed kannada-english twitter data for emotion prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6703–6709.

Bansal, S., Garimella, V., Suhane, A., Patro, J., and Mukherjee, A. (2020). Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1018–1023, Online, July. Association for Computational

Linguistics.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Carvalho, J. and Plastino, A. (2021). On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review*, 54(3):1887–1936.

Cateni, S., Colla, V., and Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135:32–41.

Chinchor, N. (1992). Muc-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, page 22–29, USA. Association for Computational Linguistics.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cortis, K. and Davis, B. (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.

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

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017). SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada, August. Association for Computational Linguistics.

Cortis, K., Verma, K., and Davis, B. (2021). Finetuning neural language models for multidimensional opinion mining of english-maltese social data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 309–314.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gupta, U., Chatterjee, A., Srikanth, R., and Agrawal, P. (2017). A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*.

Hemmatian, F. and Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52(3):1495–1545.

Hilts, A. and Yu, E. (2010). Modeling social media support for the elicitation of citizen opinion. In *Proceedings of the International Workshop on Modeling Social Media*, pages 1–4.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kothari, A., Magdy, W., Darwish, K., Mourad, A., and Taei, A. (2013). Detecting comments on news articles in microblogs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a backpropagation network. In *Advances in neural information processing systems*, pages 396–404.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer.

Ligthart, A., Catal, C., and Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, pages 1–57.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

Mamta, Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., and Saha, T. (2020). Multi-domain tweet corpora for sentiment analysis: Resource creation and evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5046–5054, Marseille, France, May. European Language Resources Association.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Padurariu, C. and Breaban, M. E. (2019). Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2):1–135, January.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das,

A. (2020). Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46.

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

Salton, G. and McGill, M. J. (1986). Introduction to modern information retrieval.

scikit learn. a). Bernoulli naïve bayes. `https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html`.

scikit learn. b). Complement naïve bayes. `https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.ComplementNB.html`.

scikit learn. c). Decision tree. `https://scikit-learn.org/stable/modules/tree.html`.

scikit learn. d). Linear support vector classification. `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`.

scikit learn. e). Logistic regression. `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`.

scikit learn. f). Nu-support support vector classification. `https://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html`.

scikit learn. g). Random forest. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`.

scikit learn. h). Support vector classification. `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yimam, S. M., Alemayehu, H. M., Ayele, A., and Biemann, C. (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.

Yu, H.-F., Huang, F.-L., and Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75.