Tagalog Sentiment Analysis Using Deep Learning Approach With Backward Slang Inclusion

Aaron John V. Boquiren, Raymond A. Garcia, Chrisrenee Jerard D. Hungria, and Joel C. de Goma
School of Information Technology
Mapua University
Makati, Metro Manila
aaronjohnboquiren@gmail.com, ragarcia@mymail.mapua.edu.ph,
chrisrenee.h123@gmail.com, jcdegoma@mapua.edu.ph

Abstract

As the internet users in the Philippines have become increasingly rampant, Tagalog Sentiment Analysis is essential when determining general sentiments correlated to a topic. Among Tagalog tweets, backward slang has gained increased popularity on social media. Deep learning algorithms such as Long-Short Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) have been proven more effective than traditional machine learning algorithms. This study explored and evaluated the effects of backward slang on two deep learning models using performance metrics for classification models. Results show that backward slang helps in the expressivity of textual data, given the improvement in overall sentiment scores.

Keywords
Sentiment Analysis, Deep Learning, Backward Slang, Netspeak, Random Search

1. Introduction
The Tagalog dialect is one of the most spoken language forms in the Philippines, wherein it can be used for communication among people orally or virtually. This dialect is also known to have specific jargons or slang that some people would use as a substitute for particular words that would possess the same meaning yet have different word structures or pronunciations. This chapter provides an overview of the research conducted, including general background and general motivation. Key points such as Sentiment Analysis are defined into specific ideas applied in the latter parts of the study.

1.1 Background of the Study
The Philippines is one of the most significant internet users as it is very rampant in the country to use technology for everyday use. As years pass in the Philippines, the population of internet users is increasing, primarily due to the pandemic's current situation, which requires the majority of the people to use these platforms to their advantage. Recently, the demand for Internet usage on jobs and education has increased the number of people using the Internet and its properties. The population of internet users (Datareportal, 2020) peaked at 73.91 million in January 2021, increasing by 4.2 million from 2020-to 2021.

As the English language forms quickly change over time, digital technology and its development allow the Philippine slang to create newer terms further and spread quickly, passing from Millennials to future digital natives (Danao et al 2017). With the increase of development on the Internet, information grows exponentially. For organizations and enterprises, extracting valuable information from consumers is very important. Social media platforms such as Facebook and YouTube collect vast amounts of user reviews that form a rich source of information for companies to understand their customers (Moudjari et al, 2020). The information provided through user reviews is used to classify sentiments essential to the decision-making process of enterprises and organizations (Li et al 2017). Li et al, 2017 explored how Chinese text processing affects sentiment analysis results. There are many differences between Chinese and English text pre-processing, such as Chinese needs to segment, while English does not need to. A total of eight kinds of pre-processing techniques were used: removing URLs, removing the numbers, removing the punctuations, removing the stop words, replacing the network words, reverting the

repeated words, replacing the emoticons, and normalizing the Chinese and English mixed texts. The results showed that Chinese and English pre-processing effects have some similarities, although the stop words in Chinese texts contain more valuable information than English texts.

The study of (Caparas et al. 2017) investigates the communicative aspects of Multilingual Code-Switching in Filipino Netspeak. The findings in the study have shown that Code Switching is mainly used by Filipino college students and industry professionals when communicating using social media. The reason why younger people prefer Code Switching. The research article (Danao et al 2017) identifies a specific type of slang in the Tagalog language: "binaliktad," or backward speech wherein syllables or characters are reversed. For example, the word "Father" is transformed into "pater" and then reversed into "erpat." According to this study, the Tagalog language game "binaliktad" is usually used by Filipinos to disguise this textual language in daytime conversations to amuse them. Also, this slang type is often used for joking and teasing to gain prestige among their peers thus, conveying positive sentiments. Furthermore, some backward Filipino texts are being used on social media for speech disguise, jokes/teasing, and slang that weaken the impact of certain expressions that might be considered improper or obscene, which affects the sentiment of a Tweet or Facebook comment present.

From the pre-processing methods of (Ilao et al, 2020), after removing unwanted expressions such as URLs, punctuations, emoticons, special characters, and apostrophes, so-called irrelevant words are removed (example: powh, jejeje, xD). These are slang words, stop words and expressions.

According to (Cirqueira et al, 2018), emoticons and slang are essential things to focus on to improve lexicon-based sentiment classification. The emoticons are the simplest way to show emotions, especially on social media sites like Twitter. Like emoticons, slang is also every day in social media, and it is needed for more accurate sentiment analysis.

The works of (Pippin et al 2015) included the emotions of Filipino Tweets in their word bank. The output of the study obtained 70% accuracy based on their performance evaluation. Based on gathered Accuracy values, the study recommended the addition of emoticons and different slangs in a word bank to improve the classification of Tagalog sentiments. This is contrary to pre-processing methods of (Ilao et al. 2020), wherein slang words and expressions are removed, affecting the reliability of their model. Also, the study (Pippin et al. 2015) noted that Ekman's six basic emotions, happiness, sadness, anger, disgust, surprise, and fear, are related to classifying sentiments on textual data.

1.2 The Problem, Gap, and Opportunity

From previous Tagalog SA research approaches (Contreras et al. 2018, Pippin et al. 2015 and Gimenez et al. 2017), expressions and slang were removed or not added to the Tagalog lexicon, which affects the reliability of identifying sentiments of Tagalog text since expressions and slangs are related to sentiment identification (Danao et al, 2017). The existing studies regarding Tagalog SA lack the enhancement of reliability and performance due to a lack of data regarding the inclusion of emoticons and chat terms/slang in the word bank. This provides an opportunity for this study to include backward slang as a feature that most Filipinos use on social media. This feature will further improve the reliability and Accuracy of Tagalog SA, thereby reducing the number of misclassifications on Tagalog sentiments.

1.3 Statement of the Problem

This research aims to utilize the Deep Learning Approach to develop a model that would classify the sentiments of Tagalog dialect texts with backward slang inclusions and identify the difference between the inclusion and neglect of backward slang to developed models. More specifically, this research seeks to answer the following questions: How would the inclusion of backward slangs affect the developed model's performance? How well would Bi-LSTM compare with LSTM in terms of performance measures? Furthermore, how well does each model correctly classify sentiments based on its features?

1.4 Objectives

The primary objective of this research is to develop a Tagalog Sentiment Analysis model that would classify the sentiments of tweets with the inclusion of backward slang. Moreover, this study shows that backward slangs are an essential feature in Twitter sentiment analysis. This research intends to do the following objectives: To evaluate the performance of LSTM to state-of-the-art Bi-LSTM with the suitable performance measures, to determine the effect

of the inclusion of backward slangs to the model's performance, and lastly, to determine what are the features that caused misclassification of sentiments from tweets.

2. Review of Related Literature

2.1 Tagalog Corpus

The researchers gathered approximately 100 articles, around 11,200 sentences to create a baseline dataset. The selected annotators used the formulated guidelines as a guide in the manual annotation. This study indicates that manual classification of sentiments may be subject to biases due to human intervention. The selection of sentiment was based on majority votes whenever there were cases where disagreement existed between the group of annotators. Labeling is translated using a Fleiss Kappa measure to account for the variation.

2.2 Preprocessing

In (Li et al. 2017), eight pre-processing operations were utilized: removing links, removing numbers, removing punctuations, removing stop words, replacing network words, reverting repeated words, replacing emoticons, and normalizing mixed Chinese and English text. Non-Arabic words and letters, punctuations, Arabic stop words, repeated words, and noise were removed as part of pre-processing steps in the study (Alnawas et al. 2019). Also, the Normalization of Arabic letters will be performed after removing the punctuations and noise to allow processing to proceed uniformly. The study (Alotaibi et al. 2019) pre-processed their data by cleaning tweets, formatting, and storing in its appropriate dataset to ensure its readiness for analysis. An appropriate normalization library removed emojis, symbols, and meaningless characters. The study (Alahmary et al. 2019) pre-processed their data by first removing duplicated tweets then applying data annotation using crowdsourcing to classify tweets into two classes, comprising 17,707 positive and 14,356 negative tweets.

2.3 Data Annotation

The study of (Abisado et al. 2018) worked on students' engagement during examinations through emotion evaluation. For the data pre-processing used by the study's authors, A licensed behavioral psychologist is involved in their study as a domain expert. The psychologist participated in the study by annotating video examples to their classification. A licensed psychologist trained three annotators to label videos with correct annotations. The labeled videos were then used for the validation of data. Kappa statistics were used to validate data annotation, a statistical measure to measure the agreement between the three video raters.

2.4 Utilization of Hashtags

While the study of (Malouf et al. 2008) implied that text-based classification performed poorly on sentiment identification on political texts without their political affiliation, the study conducted by (Alfina et al. 2017) shows that political tweets' increasing complexity due to their subjectivity to political party affiliation political viewpoint. Hashtags were considered very helpful in identifying sentiments, as the presence of these features on tweets can already provide the political viewpoint. As demonstrated in the mentioned study, exploiting hashtags on political tweets makes it crucial to classify the sentiment polarities obtaining high Accuracy and cheaper when annotating the dataset.

2.5 Deep Learning

The works of (Soufan, 2019) applied Deep Learning to classify sentiments on Arabic Textual data. The authors of this study noted the problem with feature representation, and for the Arabic language that was the object of the study, those features are costly and require much pre-processing. The dataset used in this study were gathered from Open-Sourced datasets found on the Internet. The amount of data gathered from different datasets in this study averaged about 4000 textual data. The study used traditional Machine Learning Algorithms and compared them to Deep Learning Approach. For a relatively small dataset, supervised learning methods used in this study, such as SVM and NB, outperformed Deep Learning methods such as LSTM and CNN. The study suggests that to build a Deep Learning Model for Dialectal sentiment analysis, a large amount of data is needed to train the deep learning methods effectively. The classical machine learning techniques need data to be interpreted by attributes before algorithm application. This shows a disadvantage of Feature Engineering, such as information loss and human dependence factor (Alahmary et al. 2019).

Several research studies of SA use Deep Learning concatenate on learning vectors as features without feature engineering. Existing deep learning methods for sentiment classification typically include two stages. In the first

stage, they learn word embeddings from a text corpus. In the second stage, word embeddings are applied to producing the representations of sentences/documents with semantic composition (Tang et al. 2015).

Deep learning is a subfield of machine learning where algorithms inspired by the human brain's function are called artificial neural networks to learn from large amounts of data. It is said to be better than standard neural networks since deep learning has a lot of hidden layers and could be both utilized in supervised and unsupervised learning approaches (Alahmary et al. 2019). In text classification, deep learning has primarily been used to aid feature extraction before training classifiers (Prusa et al. 2016).

This study (Alahmary et al. 2019) used Deep Learning approach to classify sentiments on Saudi Dialect and was compared to SVM to analyze and compare performance evaluation. The SA classification algorithm used in this study was SVM, Long-Short Term Memory (LSTM), and Bi-LSTM. This study found out that Bi-LSTM outperforms both LSTM and SVM in terms of Accuracy by obtaining 94% accuracy while 86.4% accuracy for SVM models.

2.6 Hyperparameter Tuning and Cross-Validation

According to (Elgeldawi et al. 2021), hyperparameter tuning captures the current performance of an existing model and compares its performance with previous models. In many machine learning algorithms, hyperparameter tuning is an essential method to obtain the optimal parameters of an existing algorithm. To perform hyperparameter tuning in a machine learning model, a set of parameters must be implemented on a model wherein the values are set before starting the process. The study (Priyadarshini et al. 2021) mentioned two standard hyperparameter optimization algorithms: Grid Search and Random Search. Whereas in grid search, every combination of values set on the parameters is being trained to obtain the best model through cross-validation, on the other hand, random search model training is based on the combination of random parameters to obtain the best performance also by cross-validation. The difference between the two methods is the time complexity of grid search as its runtime increases proportionally with the number of parameters set in the model.

In contrast, random search combines parameters based on sampled search space and evaluated sets from a specified probability distribution (Elgeldawi et al. 2021). Also, from the same study, the difference between the Accuracy on both search algorithms had a slight difference on every machine learning algorithm used. As such, the study (Ramadhani et al. 2016) recommended using random search when using hyperparameter tuning as it would have a shorter computation time and be more efficient than using grid search.

3. Methods

3.1 Conceptual Framework

Raw data will be gathered on Twitter which then will be annotated by three domain experts. The annotated dataset will undergo pre-processing to remove noise in the data further. LSTM and Bi-LSTM models will be developed in which the default set of parameters is used. Both models will undergo hyperparameter tuning to determine the best set of parameters on both models, with the main difference that two different versions of train sets are used wherein the first train set contains backward slang while the other does not. The best parameters will be validated using a test set with backward slang. After model validation on four different models, each will be validated based on performance metrics.

3.2 Tagalog Corpus

The characteristics of Tagalog Corpus would include positive and negative word data. The collected sentences will provide their respective sentiment polarity, and the partial goal of this study is to extract valuable information from these processed sentences to determine the sentiment of the new sentences.

3.3 Data Gathering

The study (Abdelli et al. 2019) recommends that there should be more data for a better result in the Deep Learning approach. With the following suggestion of additional tweets to be mined, the researchers gathered 32,036 tweets to be labeled with their respective sentiment polarity: 0 - for negative and 1 - for positive.

The initial phase of data gathering will be tweet extraction using keywords of different topics preferably, words inclusive of social trends in the Philippines. "@" tags will also be used as keywords as they may extract more

opinionated data. The extracted tweets will have a maximum character count of 280 containing text, hyperlinks, different symbols, emoticons, and short words. Raw data gathered from Twitter will be pre-processed to have better output in this study. Using the Twitter platform to collect tweets, the researchers have collected tweets containing backward slang alongside Filipino politics.

The initial tweets gathered utilizing mining the textual data amounted to 32,036 tweets. However, due to the number of neutral sentiment polarities that are needed to be removed from data annotation, only 10,307 tweets remained 5,205 belonging to positive sentiment and the other 5,101 for negative sentiment.

3.4 Data Annotation

For the dataset to become coherent and complete, assistance from three different domain experts is needed to become consistent with the model output—clear rules when labeling tweets are required to achieve good results on annotated tweets. The data will be labeled according to the criteria made by the domain experts. With the subjectivity of tweets, especially on politics are being involved in the study to ensure that majority voting is present and will be computed using kappa statistics (Abisado et al, 2018). Kappa statistics will be used to measure the agreement of the annotators to test the interrater reliability or, in other words, how the data is precisely annotated.

3.5 Data Preprocessing

Pre-processing will be the next in the process after data gathering. Based on the recommendations of research (Pippin et al, 2015), this study will not remove emoticons and Filipino expressions, slang words, and other terms. After gathering enough sentences, the data set will undergo pre-processing through the following steps: Removal of stop words, removal of non-alphabetic and special characters, Changing uppercase to lowercase, removal of duplicates, Normalization of Taglish sentences, Normalization of Filipino slangs, and lastly, Normalization of Hashtags. Table 1 shows the samples of preprocessed tweets whereas the noise from irrelevant features were removed.

Table 1. Samples of Preprocessed Tweets

Raw Text	Preprocessed Text
sobrang obob ko pala dati? naniniwala ako sa sabi-sabi? that Marcos is a hero and Duterte is the best president? hahahaha pota #MarcosNotAHero #NeverAgain #OustDuterte	sobrang obob ko pala dati naniniwala ako sa sabi sabi that marcos is a hero and duterte is the best president hahahaha pota marcosnotahero neveragain oustduterte
@iMthinkingPinoy @Siegreg1 NGANGA ang mga yellowtards , puro kyo porma. Akala ko si LENI ang ipapalit nyo kay PRRD? People Werpa wer na ? Hindi na uso ang tanga ngayong panahon na itoã€yung mga bayaran na lang ang natitiraã€, #AquinoNeverAgain	nganga ang mga yellowtards puro kyo porma akala ko si leni ang ipapalit nyo kay prrd people werpa wer na hindi na uso ang tanga ngayong panahon na ito yung mga bayaran na lang ang natitira aquinoneveragain

3.6 Feature Representation using

The researchers decided to use Word2Vec for this study's word embedding method. Word2Vec uses two methods which are the Continuous Bag-of-Words Model and the Continuous Skip-gram Model. The CBS algorithm predicts the current word based on the context of the words surrounding it. On the other hand, the Continuous Skip-gram algorithm predicts the context words surrounding the current word. The continuous Skip-gram Model was used in this study due to the lack of developed Tagalog Bag-of-Words. To ensure the uniform fixed size of tweets when fitting into two deep learning models used in this study, the dimensions are set to 300, and vocabulary size values are used in the embedding layers, which will be used for machine learning models. This method ensures that all sequences will have the same length. For the embedding layer used in this study, the weights used in this layer came from the embedding matrix based on the Word2Vec Model, and the input length has the same value as the padded length of words which is 300.

3.7 Hierarchical Softmax vs. Negative Sampling

Hierarchical Softmax (HS) is a computing method whereas it builds a tree from the advanced vocabulary of words, and the leaf nodes that represent infrequent words would inherit the previous vector representations in the tree,

which implies that it can be affected by other frequent words in the corpus. With it, a given sample must be evaluated in O(log(N)) network units, not O(N). While HS provides the elimination of preference on frequent categories, Negative Sampling (NS) is developed based on contrastive estimation, and it randomly samples infrequent words. Similar to stochastic gradient descent, it looks into negative training samples. The difference between both models is that NS causes each training sample to update only a small percentage of a model's weights. Also, it is essential to note that the computational cost for training on NS is linearly dependent on the number of noise words on each step.

To determine if this study will use either HS or NS, the output of both models was compared to the proximity of the selected words to each other. From both versions of datasets, two models developed from HS and NS were developed.

3.9 Long-Short Term Memory

LTSM has been invented as a Recurrent Neural Network (RNN) architecture designed to handle long time-dependencies. This architecture prevents RNN from learning long-term dependencies described as vanishing or exploding gradient problems. LSTM avoids this by breaking down input sequences during movement around the hidden states, hidden layers in the network. Memory units are also added to LSTM to solve the RNN problem further. This study proposes using Long-Short Term Memory (LSTM) to apply Tagalog Sentiment Analysis on the collected dataset and compare the experimental results with Bi-LSTM.

3.10 Bidirectional Long-Short Term Memory

From the proposed model of (Elfaik et al. 2020) when it comes to NLP, LSTM cannot take into account post word information because the sentence is read-only in a forward direction, which lacks the full use of contextual information. In order to solve this problem, Bi-directional LSTM is used, whereas two LSTM outputs are stacked together, one that reads in a forward direction and another for the backward direction. The hidden states of both LSTM output forward, backward, and then concatenated, producing their final hidden state.

3.12 Hyperparameter Tuning and Evaluation

To select the best parameters for both LSTM and Bi-LSTM models, hyperparameter tuning is used by creating a set of parameters to identify what model performs best in terms of Accuracy and F1- score. In this study, the researchers used random search (Elgeldawi et al. 2021) to identify optimal parameters for both models due to their efficiency and shorter computation time.

The 10,307 mined and processed tweets were used to evaluate both model performances. On both with backward slang and without backward slang data. The train set used 80% of the original dataset used for model training. The train set was further split into train and validation sets, whereas 50% of the remaining set was used for the validation set as it will be used to train the model with stratified 10-fold cross-validation, whereas nine sets were used for training, and 1 set was used for validation. The other 50% from the remaining dataset contains the test set wherein both models were evaluated. For this remaining 50%, on both models developed, the data here contains backward slang for both with and without backward slang to evaluate their performance metrics.

3.13 The 5x2 cross-validated paired t-test

To test if the difference between the Accuracy of models with backward slang and models without backward slang are statistically significant, a 5x2 cross-validation paired t-test is utilized. In contrast, a p-value is computed to identify a statistical significance that both models perform differently.

The paired t-test was introduced by Dietterich (Dietterich 1998), whereas the two models' performance was compared using this procedure. The procedure works because a dataset is split into training and a test set. In the 5x2 paired t-test, the splitting was repeated by five times. Two models being compared in the process are evaluated using the test set, obtaining their Accuracy and Receiver Operating Characteristic values. After which, the train and test sets are rotated, and the performance will be computed again, as shown in Figure 3, which will result in two different performance metrics.

In the case of this study, models without backward slang and models with backward slang will be compared to extensively identify if the difference between Accuracy and roc/AUC is statistically significant or not.

4. Results and Discussion

4.1 Result Summary

The models developed from LSTM and Bi-LSTM produced good to excellent results based on performance metrics used. Looking at each of the metrics in the results, versions of both models wherein Backward Slang was removed show a reduced performance compared to models with backward slang inclusion. Classification accuracy, precision, recall, F1-Score, confusion matrix, and receiver operating characteristic was used to evaluate each model's performance. The best model produced in this study was Bi-LSTM with Backward Slang, followed by LSTM with Backward Slang by a small margin in terms of all metrics used.

4.2 Hierarchical Softmax vs. Negative Sampling

Both Hierarchical Softmax and negative sampling were utilized in this study to compare their performance based on their approximate prediction. Table 3 and Table 4 shows the difference between model predictions on both models.
Looking at similarities between HS and negative sampling, it is generally shown that negative sampling provides better word similarities on selected words, as negative sampling shows better performance when it comes to model training from infrequent words. Table 3 shows the difference between HS and NS. It is shown that words present on word similarities are more relevant than HS on NS.

It is also evident from the corpus itself that there are more frequent words present in the vocabulary, hence the better performance of NS when obtaining its output and performance, contrary to HS. An example of the difference in classification between both similarities are from the word “Marcos,” as NS shows closer similarity and relevance to the word; “Bongbong,” which is his nickname, “jr” as he is the son of Ferdinand Marcos, “bbm” to abbreviate his nickname, and “magnanakaw,” a relative issue with the target word as the corpus itself was mined from political tweets. Contrary to HS, it shows weaker relevance and similarity to the target word. Because of better representation of outputs from NS, it will be utilized for the word2vec model and used for LSTM and Bi-LSTM model training (Table 2 and table 3)

Table 2. Word Similarities on Models Trained Without Backward Slang

Word	Euclidian Similarities							
	Hierarchical Softmax				Negative Sampling			
Country	Philippines	Better	Best	Pride	Great	Doing	father	our
Politics	Shit	Malinaw	Moreno	Essential	Real	political	pandemic	did
Marcos	kabataan	sasabihin	fans	sabi	bongbong	jr	bbm	magnanakaw
China	tsina	dagat	covid	gobyerno	tsina	virus	dagat	pinas
Leni	madam	kalaban	du30	bbm	lugaw	madam	kiko	dapatsileni

Table 3. Word Similarities on Models Trained with Backward Slang

Word	Euclidian Similarities							
	Hierarchical Softmax				Negative Sampling			
Country	Philippines	Father	part	better	father	great	better	doing
Politics	malinaw	mass	pang	Pacquiao	father	great	better	doing
Marcos	BBM	ayon	raw	fans	bongbong	BBM	jr	protect marcosjr
China	tsina	government	secretary	dagat	tsina	line	virus	dagat
Leni	campaign	manny	BBM	ping	madam	lugaw	letlenilead2022	campaign

4.3 Hyperparameter Optimization

The initialization of the model was initiated wherein the values will prevent excessive overfitting and underfitting. To achieve optimal parameters for the developed model, hyperparameter tuning using random search was performed, including the number of units, LSTM/Bi-LSTM dropout, dropout rate, learning rate, batch size, and the number of epochs. A model checkpoint has been developed to monitor the lowest validation loss to save the model's weights on each run. Using random parameter combinations, the dropout rate was set from 0.2, 0.3, and 0.5. For the learning rate, 0.0001 was selected. For batch size, 32 and 256 were selected with a spacing of 32—furthermore, 50 to 256 for epochs with a selected value of 10.

For table 4 and table 5, the LSTM model trained without backward slang obtained a mean accuracy of 89.22%, having the number of LSTM units by 90 and 10, dropouts of 0.2, batch size of 64, and 50 epochs while Bi-LSTM achieved a value of 89.03% accuracy having the number of units of 120 and 220, dropout of 0.2 and 0.3, the value of learn rate being the same with LSTM which is 0.0001, batch size of 96, and epochs of 250.

Table 6 shows the following optimal hyperparameter combinations on models trained with backward slang. The mean accuracy of the optimal parameters on LSTM achieved 92.18% which combined units of 150 and 80, dropout rate of 0.2, batch size of 64, and 80 epochs. For the Bi-LSTM, the units are 240 and 150, dropout rate of LSTM layer is at 0.2, dropout layer rate is at 0.5, 150 number of units, batch size of 32, and epochs of 100, obtaining a mean accuracy of 92.16% (Table 5).

Table 4. Best Parameter Combination for models trained without backward slang

LSTMNoBWSl		Bi-LSTMNoBWSl	
Mean Accuracy	89.22 %	Mean Accuracy	89.03 %
LSTM Units	90	LSTM Units	120
LSTM Dropout	0.2	LSTM Dropout	0.2
Dropout Rate	0.2	Dropout Rate	0.3
Units	10	Units	200
Learn Rate	0.0001	Learn Rate	0.0001
Batch Size	64	Batch Size	96
Epochs	50	Epochs	250

Table 5. Best Parameter Combination for models trained with backward slang

LSTMNoBWSl		Bi-LSTMNoBWSl	
Mean Accuracy	92.18%	Mean Accuracy	92.16%
LSTM Units	150	LSTM Units	240
LSTM Dropout	0.2	LSTM Dropout	0.2
Dropout Rate	0.2	Dropout Rate	0.5
Units	80	Units	150
Learn Rate	0.0001	Learn Rate	0.0001
Batch Size	64	Batch Size	32
Epochs	80	Epochs	100

4.6 Long-Short Term Memory

Starting with the examination of Table 7, 3.82% difference in Accuracy and AUC was computed, indicating improvement in the Accuracy of LSTM with the inclusion of Backward Slang. Other performance metrics have been computed. However, it shows that the remaining metrics have indicated improvement, most notably, recall score, accumulating for 5.14%, as the model LSTM in which Backward Slang was included increased 4% compared to LSTM without Backward slang. In comparison, the presence of performance metrics and their differences were measured. The crucial metric that must be considered when indicating the performance of a model is classification accuracy and F1-score, wherein it measures the test accuracy by finding the balance between precision and recall (table 6).

Table 6. Performance Metrics on LSTM

	LSTMNoBWSl	LSTMWithBWSl	Difference
Accuracy	87.43 %	91.71 %	4.28 %
Precision	87.99 %	91.51 %	3.52 %
Recall	86.71 %	91.95 %	5.24 %
F1-Score	87.34 %	91.73 %	4.39 %
ROC/AUC	87.43 %	91.71 %	4.28 %

4.9 Bidirectional LSTM

After computing performance metrics on Bi-LSTM, Table 8 shows the performance metrics on validated models of Bi-LSTM with and without Backward Slang. One notable difference is the values of metrics in Bi-LSTM contrary to LSTM, as shown in Table 7. It is shown that Bi-LSTM obtained better performance compared to models developed using LSTM, most notably, on recall score. The same can be said from what was that classification accuracy, and F1-score are both important when looking at the performance metrics of a model. Upon investigation on the values in Table 8, the difference between classification accuracy and AUC on both versions of Bi-LSTM has accumulated to 3.45% difference while F1-Score obtained 3.56% difference between models. A lower value in difference between the two models regarding Accuracy and F1-score could mean that Bi-LSTM has played a role in minimizing the misclassification of labeled tweets.

Table 7. Performance Metrics on Bi-LSTM

	LSTMNoBWSl	LSTMWithBWSl	Difference
Accuracy	88.31 %	91.76 %	3.45 %
Precision	88.88 %	91.43 %	2.55 %
Recall	87.58 %	92.14 %	4.56 %
F1-Score	88.23 %	91.79 %	3.56 %
ROC/AUC	88.31 %	91.76 %	3.45 %

6. Conclusion

This study conducted a series of experiments wherein the Backward Slang was included on tweets as a feature to classify sentiment polarities. The models developed in this study do not use any feature engineering to extract special features or any complex modules like sentiment treebanks. Every version of models in this study only relies on pre-trained word vector representation. The study used LSTM and Bi-LSTM as these deep learning architectures can extract further contextual information from dealing with forward and backward dependencies from feature sequences. Despite the complexity of Filipino tweets, especially in the domain of politics due to the presence of sarcastic and satire tweets, the models developed can classify their respective sentiment polarities, having good to great performance metrics. It is conclusive that the inclusion of Backward Slang helps improve overall sentiment scores and performance metrics. While Backward Slang can be used to express negative or positive opinions on a tweet, the presence of these features on tweets has appeared to improve the expressivity of negative opinions more than positive opinions on sentiment analysis.

For future studies, the inclusion of another domain aside from politics when mining tweets is encouraged to be explored as the study is limited to identifying political tweets wherein the presence of sarcastic and satire tweets may often occur in the corpus.

References

Ang Li and Yunfang Chen. Pre-processing Analysis for Chinese Text Sentiment Analysis. In Proceedings of the 2017 2nd International Conference on Communication and Information Systems (ICCIS 2017). Association for Computing Machinery, New York, NY, USA, 318–323, 2017.

Anwar Alnawas and Nursal Arici. Sentiment Analysis of Iraqi Arabic Dialect on Facebook Based on Distributed Representations of Documents. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*. 18, 3, Article 20 (July 2019), 17 pages, 2019.

Abdelli, F. Guerrouf, O. Tibermacine and B. Abdelli,, Sentiment Analysis of Arabic Algerian Dialect Using a Supervised Method," *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), Taza, Morocco,* pp. 1-6, 2019.

Ayah Soufan. Deep Learning for Sentiment Analysis of Arabic Text. *In Proceedings of the ArabWIC 6th Annual International Conference Research Track (ArabWIC 2019). Association for Computing Machinery, New York*, NY, USA, Article 20, 1–8, 2019.

Caparas, Pilar & Gustilo, Leah., Communicative aspects of multilingual code switching in computer-mediated communication. 2017 *Indonesian Journal of Applied Linguistics*. 7. 2011.

Derra, N. D., & Baier, D., Working in detail: How LSTM hyperparameter selection influence... 1970 Archives of Data Science, Series A. Retrieved February 8, 2022. 1970, January 1.

Dietterich TG, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput 10*:1895–1923, 1998.

Digital 2021: THE PHILIPPINES. 11 February, https://datareportal.com/reports/digital-2021-philippines#:~:text=There%20were%2089.00%20million%20social,total%20population%20in%20January%202021. Accessed: 2021-04-08, 2021.

D. Cirqueira, M. Fontes Pinheiro, A. Jacob, F. Lobato and Á. Santana, A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media, IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, 2018, pp. 746-749, 2018.

Elfaik, Hanane & Nfaoui, El Habib. Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text. Journal of Intelligent Systems. 30. 395-412. 10.1515/jisys-2020-0021, 2020.

Elgeldawi, E., Sayed, A., Galal, A.R., & Zaki, A.M., Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. *2021 Informatics,* 2021.

G. Danao, J. Torres, J. V. Tubio and L. Vea, Tagalog regional accent classification in the Philippines, *2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Manila, Philippines* pp. 1-6, 2017.

Ika Alfina, Dinda Sigmawaty, Fitriasari Nurhidayati, and Achmad Nizar Hidayanto., Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain. *2017 In Proceedings of the 9th International Conference on Machine Learning and Computing (ICMLC 2017). Association for Computing Machinery, New York, NY, USA*, 43–47, 2017.

Ilao, Adomar & Fajardo, Arnel. (2020). SENTIPUBLIKO: Sentiment Analysis of Repost Jejemon Messages using Hybrid Approach Algorithm. IOP Conference Series: *2020 Materials Science and Engineering*. 938. 012010. , 2020.

Jennifer O. Contreras, Melvin A. Ballera, Ace C. Lagman, and Jennalyn G. Raviz. 2018. Lexicon-based Sentiment Analysis with Pattern Matching Application using Regular Expression in Automata. *2018 In Proceedings of the 6th International Conference on Information Technology: IoT and Smart City (ICIT 2018). Association for Computing Machinery, New York, NY, USA*, 31–36, 2018.

Jr, Pippin, & Odasco, Ron & Jr, De & Tolentino, Miguel & Bringula, Rex., Classifications of Emotion Expressed by Filipinos through Tweets. *2015 Lecture Notes in Engineering and Computer Science*. 1. 292-296, 2015.

Kumar Ravi, Vadlamani Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, *2015 Knowledge-Based Systems,* Volume 89, 2015, Pages 14-46, ISSN 0950-7051, 2015.

Leila MOUDJARI, Karima AKLI-ASTOUATI. An Experimental Study on Sentiment Classification of Algerian Dialect Texts, *Procedia Computer Science, Volume 176,* Pages 1151-1159, ISSN 1877-0509, 2020.

L. Cheng and S. Tsai, "Deep Learning for Automated Sentiment Analysis of Social Media, *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada*, pp. 1001-1004, 2019.

Malouf, Robert & Mullen, Tony. (2008). Taking sides: User classification for informal online political discourse. *2008 Internet Research*. 18. 177-190, 2008.

Monderin, C. and Go , M.B.. Emerging Netspeak Word Choices in Social Media on Filipino Pop Culture. *2021 International Journal of Linguistics, Literature and Translation*. 4, 6 (Jun.2021), 49-61, 2021.

M. Abisado, B. Gerardo, L. Vea and R. Medina, "Experimental Facial Expression and Gesture Training Towards Academic Affect Modeling," *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, Baguio City, Philippines, pp. 1-4, 2018.

Priyadarshini, I., & Cotton, C. , A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis. *2021 The Journal of Supercomputing*, 1 – 22, 2021.

RM.Alahmary, H. Z. Al-Dossari and A. Z. Emam, "Sentiment Analysis of Saudi Dialect Using Deep Learning Techniques," 2019 *International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand*, pp. 1-6, 2019.

SAlotaibi, R. Mehmood and I. Katib, "Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect," *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), Rome, Italy*, pp. 330-335, 2019.

Tang, Duyu & Qin, Bing & Liu, Ting., Deep learning for sentiment analysis: Successful approaches and future challenges. *2015 Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 5. 292-303. 10.1002/widm.1171, 2015.

Biography/Biographies

Aaron John V. Boquiren is an undergraduate BSCS student in Mapúa University. He is performance-driven and collaborative individual student who seeks to further hone his skills in Artificial Intelligence and Data Analytics.

Raymond A. Garcia is an undergraduate BSCS student in Mapúa University. He is an aspiring Data Scientist that seeks to further improve his skill in data analysis and understanding in Machine Learning Algorithms.

Chrisrenee Jerard D. Hungria is an undergraduate BSCS student in Mapúa University. He aims to take into his career the path wherein Artificial Intelligence, Data Analytics or Data science is involved. He is a consistent academic scholar (2018-2021) and strives to gather the most experience he can throughout his career.

Joel C. De Goma is a PhD in Computer Student of Mapúa University. He specializes in Artificial Intelligence and Data Science.