Comparative Evaluation of Lexicons in Performing Sentiment Analysis

Wan Nur Syahirah Wan Min[1] and Nur Zareen Zulkarnain[2]
[1,2] Centre for Advanced Computing Technology (C-ACT),
Fakulti Teknologi Maklumat Dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia.
Email[2]: zareen@utem.edu.my

Abstract—Twitter is one of the fastest growing social media platforms which allows users to express themselves in short text messages on a wide range of topics. The amount of text produced allows for the understanding of human behaviour. One of the analysis that can be performed is sentiment analysis. Even though sentiment analysis has been researched for many years, there are still several difficulties in performing it such as in handling internet slangs, abbreviations, and emoticons which is common in social media. This paper investigates the performance of two lexicons which are VADER and TextBlob in performing sentiment analysis on 7,997 tweets. Out of the 7,997 tweets, 300 tweets were then randomly selected and three experts in psychology and human development were asked to classify the tweets manually based on three polarities. From the study, it is found that both lexicons have an acceptable accuracy rate of 79% for VADER and 73% for TextBlob. Considering all of the performance score, VADER emerged as a better lexicon as compared to TextBlob. The result of this study serves to help researches in deciding which lexicon to use in performing sentiment analysis for social media texts including microblogs.

Index Terms— Sentiment analysis, opinion mining, lexicon, social media, Twitter.

I. Introduction

THIS growth of social network platform has brought many people fresh ways in which internet data are generated and consumed. One form of social network platform is microblogging where this platform has been dubbed the fastest growing form of digital communication [1]. Twitter is a social media network that dominates microblogging. Twitter messages, commonly called tweets, are much shorter than normal microblogging platforms. With a maximum of only 280 characters, it allows individuals to access and read information on Twitter more easily. Through Twitter, users can choose from various individuals to different organisations they would like to monitor or "follow" and would have access to data and communication on various subjects ranging from news to home purchasing [2].

With the vast amount of conversations and opinions being presented on Twitter, there are a lot of opportunities to study and understand the human behaviour on social network. Among the studies that can be performed is sentiment analysis. Broadly, there are two types of sentiment analysis approaches which are machine learning-based and lexical-based. Machine learning methods often use supervised classification techniques, which are framed as a binary (i.e. positive or negative) for sentiment identification. In this way, labelled data are required to train classifiers. According to Pennebaker, Francis and Booth [3], the limitation of using machine learning methods is that there is low availability of labelled data which then causes the low applicability on new data. This limitation could be caused by the high cost of labelling even only small data for simple tasks.

Contrastively, lexical-based methods use a predefined list of words where every term is associated with specific sentiments and polarities [4]. The lexical-based methods differ depending on their contexts in which they have been created. Even though lexical-based methods do not rely on any labelled data, the challenge in using lexical-based methods is to find a unique lexical dictionary that have been created to be used in various contexts [5]. This is especially applicable in the context of microblogging as lexical dictionary often do not cater for slangs, short forms and internet acronym which are heavily used in Twitter to stay within the character limit.

The purpose of this paper is to compare the performance of two widely used lexicons, VADER (Valence Aware Dictionary for sEntiment Reasoning) and TextBlob in performing sentiment analysis on microblogs in this case, Twitter. VADER is a lexicon that was specifically tailored for microblogs and is used to produce and empirically validate a combination of qualitative and quantitative methods to analyse the sense of a Twitter context [6]. On the other hand, TextBlob is an open source python library that reuses NLTK corpora to perform sentence level analysis of text [7] for any type of content. It also provides scores for polarity and subjectivity of each word in the text [8].

There have been several researches in comparing the performance of various lexicons in performing sentiment analysis on text [9-11]. However, these researches mainly compare the usage of general lexicons such as the General Inquirer and SentiWordNet which does not consider internet slangs and emoticons. This paper, however, is interested in comparing the performance of VADER and TextBlob in performing sentiment analysis on microblogs where in this study we have used the data from Twitter. The limitation of words in microblogs requires users to use internet slangs and acronyms as well as emoticons in order to express themselves. Therefore, since both VADER and TextBlob considers these writing styles, the study is interested to investigate which lexicon performs better. The result of this study will help in deciding which lexicon to use when performing text analysis on short social media texts.

The rest of this paper is organized as follows: Section 2 will discuss works related to using lexicons in sentiment analysis as well as a few works that compares the performance of these lexicons. This includes reviewing several existing affect lexicons including VADER and TextBlob. Section 3 will explain the experimental setup of this study while section 4 will present and discuss the result of the experiment. Finally, section 5 will conclude the study and suggest the better lexicon to be used in performing sentiment analysis for social media short texts.

II. Related Works

Sentiment analysis or opinion mining relates to the use of a method to evaluate and extract the sentiment that a writer expresses in a text. The positive, negative, and neutral classification of the text's sentiment is the result of a general assessment of sentiment by assigning the polarity and subjectivity value. The polarity indicates the positivity and negativity of the text while the subjectivity indicates the opinion or judgement of the text, whereby the closer the value to +1, the more judgemental the text is.

In essence, there are two main approaches in analysing sentiment. The first approach is by using machine learning techniques such as deep learning whereby the text are classified based on sentiment (text classification) and secondly is by using lexicon-based approaches where the text are compared with a dictionary of words to understand the sentiment (text matching) [12]. A lexicon is a collection of information about a language's word on its lexical category. It is like a dictionary of a language or subject and a total stock of words that carry meanings [13].

Basically, there are two types of approaches for using lexicons [14]. First, the dictionary-based approach whereby seed terms are usually collected and annotated manually before being grown by searching for their synonyms and antonyms in a dictionary. Second, the corpus-based approach where it is domain specific and the set of opinion terms grows by searching for related words using either statistical or semantic techniques. In this paper, we are more interested in investigating the corpus-based approach in analysing Twitter sentiments.

A. Review of Existing Sentiment Lexicon

Using sentiment lexicons in analysing texts is not something new. There have been many previous works that analyse the sentiment of texts using sentiment lexicons such as the General Inquirer [15], SentiWordNet [16], Q-WordNet [17], WordNet-Affect [18], Linguistic Inquiry and Word Count (LIWC) [2] and Affective Norms for English Words (ANEW) [19].

The General Inquirer [15] is one of the oldest lexicons that uses binary classification of sentiment using positives and negatives where it consists of 1915 positive words and 2291 negative words. On the other hand, SentiWordNet [16] is an extensively used tool in the field of opinion mining, based on a lexical English dictionary called WordNet. This lexical dictionary categorised the prepositional phrases, verbs, nouns, and other grammatical classes into synonym sets which they call synsets. SentiWordNet utilizes three types of ratings to show the sentiment of a text which are positive, objective (neutral) and negative by matching the word in the text to WordNet dictionary synset. Q-WordNet [17] is another example of lexicon that uses WordNet as the starting point in developing a lexicon that automatically annotates WordNet senses using positive and negative polarity. A text will be associated to a positive or negative connotation based on the polarity classification assigned to them. Another example of lexicon that builds upon WordNet is the WordNet-Affect [18] lexicon. This lexicon was developed by mapping the words in WordNet with a lexical database called AFFECT that consists of 1903 terms directly or indirectly referring to mental states.

Linguistic Inquiry & Word Count (LIWC) is a widely used text-analysis software that contains a set of 73 lexicons with over 2300 words [2] which enables research in text specimens of various parts of the emotions, cognitive, structural, and processes. Table I shows example of words in 5 of these 73 lexical categories. LIWC is a trusted lexicon for researches who wants to work with emotional and sentimental polarity as the lexicon has been validated by psychologists, sociologists, and linguists, both internally and externally in over a decade's work [20].

LIWC has been widely used in understanding the affective states of users in microblogs such as Twitter. For instance, Tumasjan et al. [21] uses LIWC to understand the publics' sentiment towards a political campaign based on what they shared on Twitter. Coppersmith et al. [22] on the other hand, measures post-traumatic stress disorder using Twitter data while Abbar et al. [23] uses Twitter to provide insight into the dietary choices of 210K US-based users.

TABLE I
EXAMPLE OF WORDS IN 5 OF 73 LIWC LEXICAL CATEGORIES [2]. THE * SIGNALS THAT THE WORD IS A PREFIX AND ALL WORDS WITH THE SAME PREFIX ARE PUT IN THE SAME CATEGORY

Positive Emotion	Negative Emotion	Insight	Inhibition	Negate
happy	cry	aware*	avoid*	no
appreciat*	terrify	decid*	prevent*	nowhere
great	fear	feel	safe*	never
perfect*	anger*	know	wait	nothing
terrific	griev*	notice*	wary	aren't
value	despair*	sense	limit*	without
interest	suffers	think	stop	cannot

The Affective Norms in English Words (ANEW) lexicon contains a range of mental standards for 1,034 English phrases [19]. In contrast to Linguistic Inquiry and Word Count (LIWC) or the General Inquirer, the words in ANEW are categorized according to terms of enjoyment, excitement, and domination. ANEW words have a corresponding feeling valence in between one to nine with a neutral middle of five. Any words that values less than five are considered unpleasant or negative while values higher that five are considered pleasant or positive. For example, the value for the word "betrayal" is 1.68, "bland" is 4.01 and "dream" is 6.73 which indicates that "betrayal" and "bland" are negative while "dream" is positive. The valence value over and above the simple binary of positive and negative orientations allows for the study of the intensity of feeling in microblogging.

The lexicons mentioned in this section have been widely used in sentiment analysis but have not been developed to handle social media texts with internet slangs and emoticons. Therefore, this study explores another two lexicon which are VADER and TextBlob that focuses on social media texts.

B. Comparison of Lexicons in Existing Works

There have been several works that compares the performance of lexicons in performing sentiment analysis. The comparisons however were made using lexicons that were not developed specifically for social media texts. Musto, Semeraro, and Polignano [9] in their work compares four lexicons which are SentiWordNet, SenticNet, Wordnet-Affect and MPQA. Their comparison was made based on the lexicons performances in classifying tweets in the SemEval-2013 and Stanford Twitter Sentiment (STS) datasets which contains 14,435 and 1,600,000 tweets, respectively. Based on their

study, SentiWordNet emerges best in the SemEval-2013 dataset while SenticNet performs better on the STS dataset. Their result was controversial as SenticNet was the worst performing lexicon on the SemEval-2013 dataset. It is later found that the reason SenticNet performs badly is because of its incapability to classify neutral tweets.

Another work by Khoo and Johnkhan [10] compares five existing lexicons which are Hu & Liu Opinion Lexicon, MPQA Subjectivity Lexicon, General Inquirer, NRC Word-Sentiment Association Lexicon and SO-CAL lexicon together with a general-purpose lexicon developed by them called WKWSCI. Their study compares the effectiveness of these lexicons using Amazon product review and news headlines corpus. Their result found that the effectiveness of a lexicon depends on the corpus in which they are being used. A more recent work by Bonta and Janardhan [11] compares the performance of NLTK, VADER and TextBlob in classifying the sentiment from movie reviews. A dataset consisting of 11861 sentence-level snippets from www.rotten.tomatoes.com were used in assessing the performance of the lexicons. From their study, it is found that VADER outperforms the other lexicons.

The existing comparison works mention in this section mostly use general purpose lexicons. This study, however, is more interested in assessing lexicons that were developed specifically for social media texts. The work by Bonta and Janardhan [11] can be used as a comparison with this study as it also uses VADER and TextBlob even though on movie review dataset.

C. VADER

VADER developed by Hutto and Gilbert [6] is a simple rule-based lexicon created based on the opinion of crowd knowledge. The human-centred approach gathers intensity rating on candidate lexical features from independent raters using Amazon Mechanical Turk (AMT). VADER was developed with social media style text in mind with the aim for it to be fast enough for streaming data. Basically, VADER consists of 2 types of lexicon. First, a dictionary of words with a total of 7,517 words and secondly, a dictionary of emoticons with a total of 3,570 emoticons. Each word and emoticon in the lexicon are assessed by the independent human raters and rated from -4 as “Extremely Negative” to 4 as “Extremely Positive” and 0 as having “Neutral” emotions. Since VADER was developed for social media, the lexicon also includes sentiment-related acronyms and commonly used slangs that have sentiment values.

In performing analysis, VADER does not require preprocessing, because the assessment score includes all capitalism, punctuation, and other colloquialisms, by combining a dictionary that maps lexical characteristics to emotional intensity and other easy heuristics [24]. By summarizing the intensity of each word in the text, the value of a text can be achieved. The advantage of using VADER is that the values of intensity of these colloquial are also mapped, such as "LOL" and "meh" slang. The sentiment score is calculated through a summary of each of the words’ emotion in one sentence. Even if a single word has a sentiment score of between -4 to 4, the result for a phrase is normalized to be between -1 and 1 using the normalization formula:

$$\frac{x}{\sqrt{x^2+\alpha}} \quad (1)$$

where x is the sum of the sentiments of the phrase and alpha is a standardization parameter of 15. Table II presents ten examples of words with their respective sentiment score in VADER. From the table it can be seen that VADER considers internet acronyms such as “xoxo” and “bz” and emoticons such as “:’(” which are heavily used in social media by giving sentiment scores to these words.

TABLE II
WORDS WITH SENTIMENT SCORE IN VADER LEXICON

Words	Sentiment Score
true	1.8
xoxo	3.0
love	3.2
melancholy	-1.9
nasty	-2.6
adopt	0.7
sorrow	-2.4
:'(	-2.2
bz	0.4
blame	-1.4

The sentiment score ranges from -4 to 4 but when combined as a phrase, the words “true nasty xoxo” will give the values {'neg': 0.346, 'neu': 0.0, 'pos': 0.654, 'compound': 0.4939}. The first three values are the negative, neutral, and positive values which have been normalised. The compound score is the sum of all the standardized lexicon scores from -1 to 1. In this case, the positive sentiment of the phrase is quite low with a compound value of 0.4939.

D. TextBlob

TextBlob is a python natural language processing library that contains a sentiment lexicon with 2,919 words each with a polarity and subjectivity score [7]. It contains two sentiment analysis implementations which are PatternAnalyzer and NaiveBayesAnalyzer. PatternAnalyzer was based on the Pattern library [25] while NaiveBayesAnalyzer used an NLTK classifier trained on a movie reviews corpus. Sentiment analysis using TextBlob returns a tuple of polarity and subjectivity scores where polarity is a float between -1.0 to 1.0 and subjectivity is a float in the range of 0.0 to 1.0 whereby 0.0 is highly objective and 1.0 is very subjective. Sentiment matching is implemented by calculating the polarities of a text and if the result is less than 0, it will be labelled as negative, more than 0 will be positive and 0 will be regarded as neutral. Other than affect words, TextBlob also handles modifiers, intensifiers, and negation such as “very”, “fairly”, “really” and “not”. Each of these words are also assigned with polarities and subjectivities in the lexicon.

Fig. 1 shows the difference of the polarity and subjectivity score of a word when modifiers, intensifiers and negations are introduced. The word “great” on its own has a polarity score of 0.8 and a subjectivity score of 0.75. The word “very” on the other hand has the polarity and subjectivity score of 0.2 and 0.3, respectively. When both words are combined, the phrase “very great” has an increased polarity and subjectivity score of 1 and 0.98. Both the polarity and subjectivity score were reduced when the negation word “not” is introduced. The phrase “very great” now has the polarity and subjectivity score of -0.31 and 0.58 when the word “not” is introduced.

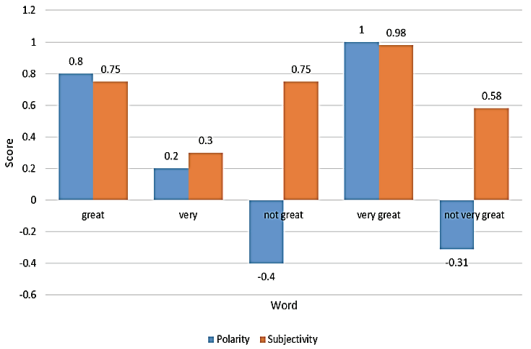

Fig. 1. Polarity and subjectivity score of words with and without modifiers, intensifiers, and negations.

When the word "very" is recognized as a modifier, TextBlob ignores polarity and subjectivity by using only the intensity to amend the words. The intensity of the word "very" is 1.3 which will be multiplied with the polarity and subjectivity of "great" to result to the new score. For the polarity score, the multiplication result will exceed 1 (0.8 when multiplied by 1.3 will become 1.04), so the polarity result is maxed to 1. For the subjectivity score, it changed from 0.75 to 0.975 after being multiplied by the intensity of the word "very". In some cases, TextBlob will ignore one-letter words or any words that it does not understand in its sentiment phrases. TextBlob eventually finds words and phrases to which polarity and subjectivity can be assigned and measures them all together for a longer text.

III. Experimental Setup

A. Data Collection

In evaluating the performance of both VADER and TextBlob in performing sentiment analysis on microblogs, a dataset of 7,997 tweets was collected from a single user from the year 2013 to 2019. The dataset consists of two columns which are the dates and the text of the tweets. This specific user was chosen for the study mainly because the person actively participates in Twitter by tweeting and retweeting every day and uses English in all of the tweets. The user profile is also set to public which allows data to be pulled from the account. Out of the 7,997 tweets collected from the user account, 4,753 tweets are original tweets from the user while 3,244 tweets were retweeted from other users. In extracting the tweets, the Twitter streaming API was used together with Tweepy, which is a Python library. Tweepy offers a convenient interface for iterating object kinds and allows up to 3,200 tweets extraction at a time. Fig. 2 shows a snippet of the tweets collected in the dataset.

	A	B
1	Date	Tweet
2	24/10/2016 14:22	@russellprue (duh typo error) part 2: social networking??.. there isn't even room for people
3	19/10/2016 18:38	my head hurts
4	17/10/2016 17:25	@James_Phelps I was sad yesterday..1 of the guys from my fave band quit the band and they replaced him
5	17/10/2016 17:23	Finally. Deleted her number from my phone contacts #heartbreak
6	16/10/2016 9:55	TWEEEEEET! good morning twitterland! going to work at 1 so need to keep packing and cleaning this flat! move ou
7	15/10/2016 7:51	@nawty I've been well, thx. Just getting up for an early morning meeting and wondering why I agreed to it. Don't li
8	14/10/2016 11:48	And this picture of her with Zack from Saved By The Bell http://twitpic.com/669lg
9	14/10/2016 6:45	VERY upset.... Mom is in the hospital
10	12/10/2016 15:38	stuck in stupid Jeuno with flag up. Wish I was in Windy fishing
11	11/10/2016 7:09	@DexterAddict Aw *hugs* Sorry man. If I were rich I'd buy u a ticket NOW!
12	11/10/2016 7:08	I will get my business coursework done by today. I will. I will.. I won't.
13	10/10/2016 12:24	Mortified. could lose my job
14	10/10/2016 12:20	@mateoviento don't see the link on my web interface
15	9/10/2016 14:09	@gfalcone601 ohhh ok. thats upsetting sorry for wasting your time.. xx
16	9/10/2016 13:32	I want 2 get my nails done today, but according 2 superstition I can't. Guess I'll have 2 wait til tomorrow then.. off
17	9/10/2016 7:25	@spotibot I am afraid you've had a bit of a #fail as the last two recommendations aren't available in the UK
18	9/10/2016 7:13	@BeckyBuckwild Omg Beckky i love you! you should've won the 250G's i was rootin for you.. hows things going?
19	9/10/2016 7:10	dunno where to start, so many things to do...
20	6/10/2016 18:19	someone please tighten some bolts in my brain :s there are too many parts loose, some might even be missing

Fig. 2. Screenshot of a few of the tweets collected from the user.

B. Data Preprocessing

In performing sentiment analysis, the data collected will often need to be preprocessed and cleaned. However, this is not necessary when using VADER since their sentiment score considers capitalism, punctuation, and other colloquialisms. TextBlob on the other hand requires the data to be preprocessed when performing sentiment analysis, however, these are performed automatically using the NLTK corpora. The first step in preprocessing the data is to reduce all words to lower case. This is a common approach whereby words such as "Me" is reduced to "me" for simplicity. While it is generally useful to decrease, this may not apply to all tasks. Proper capitalization facilitates the detection of proper nouns and increases translation accuracy. Next, stop words were removed using the nltk.corpus library. Then, tokenization is performed by cutting phrases into parts, known as tokens, by throwing away some characters such as punctuation and repeated words that could negatively affects the NLP analysis. This relies on language algorithms such as the NLTK, which are pre-trained. The RegexpTokenizer(r'\w+') function is used as a tokenizer to separate a sentence into words without punctuation.

The next step in preprocessing the collected data is to perform POS tagging of the phrases in a corpus based on its definition and context to mark a respective section of a speech tag. Once it is done, lemmatization is then performed. The aim of lemmatization is to get to the root of the word. For example, the word "attempting" will be converted to "attempt" which is the root word. Lemmatization is much more accurate compared to stemming as it uses WordNet-based strategy which effectively turns words into the real root. Twitter abbreviations and acronyms are quite complicated to be implemented for any NLP analysis. Storing all the alternatives of these abbreviations and acronyms is memory-expensive, search-delaying, and time-consuming. Therefore, it is essential to normalize certain letters automatically as a means of avoiding tool degradation and to enhance the identification of polarity. Normalization applies the standardization of text that is essential to noisy text such as commentary on twitter and text messages, when abbreviations, misprints and usage of vocabularies are common, since they can enhance the precision of sentiments. Finally, noise removal is performed by removing figures, domain-specific keys (e.g., ' RT ' for retweet) and source code. Fig. 3 shows a summary of the steps taken in the data preprocessing.

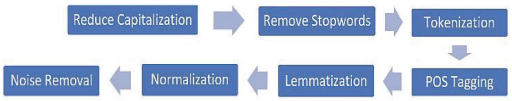

Fig. 3. Data preprocessing steps.

C. Development of Gold Standard

After the above-mentioned preprocessing steps have been applied, all 7,997 tweets collected will be analysed and classified by both VADER and TextBlob lexicons as either positive, negative, or neutral sentiment. To assess the accuracy of the classification, a gold standard was developed whereby three experts in psychology and human development were asked to classify 300 random tweets out of the 7,997 tweets analysed. The classification from the experts were performed separately and the majority decision of each tweet will be selected as the final classification that will be used in the gold standard to compare with results from VADER and TextBlob.

IV. Result and Findings

In assessing the performance of both lexicons as compared to the gold standard, the performance metrics used were accuracy, precision, recall and F-measure using the following formulas:

$$Accuracy = \frac{number\ of correctly\ classified\ positive, neutral\ and\ negative\ tweets}{number\ of\ positive, neutral\ and\ negative\ tweets\ in\ gold\ standard} \quad (2)$$

$$Precision = \frac{number\ of correctly\ classified\ positive, neutral\ or\ negative\ tweets}{number\ of\ positive, neutral\ or\ negative\ tweets\ classified\ by\ lexicon} \quad (3)$$

$$Recall = \frac{number\ of correctly\ classified\ positive, neutral\ or\ negative\ tweets}{number\ of\ positive, neutral\ or\ negative\ tweets\ in\ gold\ standard} \quad (4)$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

The 300 random tweets classified by the experts were submitted to both VADER and Text Blob. The results were then recorded and compared with the gold standard. If the lexicon and the gold standard both classify a tweet as positive, neutral, or negative, it will be labelled as PP, NENE and NN, respectively. However, if the lexicon classifies a tweet as neutral or negative when the gold standard is positive, it will be labelled as PNE or PN, respectively. A tweet will be labelled as NEP or NEN if the lexicon labels it as positive or negative when it is neutral in the gold standard. Finally, if the tweet is classified as positive or neutral when the gold standard classifies it as negative, it will be labelled as NP or NNE, respectively. Fig. 4 shows an excerpt of the labelling of the tweets.

The results of the labelling process are then compiled and summarized in a confusion matrix as presented in Table III and IV. A confusion matrix presents the predicted classification as given by VADER and TextBlob over the actual classification given by the experts in the gold standard. The confusion matrix is then used to calculate the accuracy, precision, recall and F-measure of both lexicons.

Tweet	VADER	TextBlob	Gold Standa	C (VADER	C (TextBlob
Have been rolling on the bed for the	negative	negative	negative	NN	NN
Good morning tweeps. Busy this a.m.	negative	negative	negative	NN	NN
I had a bad dream an now i cant go t	negative	negative	negative	NN	NN
I can't sleep without him nearby	negative	neutral	negative	NN	NEN
Omg, I'm a murderer! I accidently just	negative	negative	negative	NN	NN
Aww.. I lost 3 followers. FOLLOW ME	negative	negative	negative	NN	NN
is doing boring accounts stuff	neutral	negative	negative	NEN	NN
apparently even Novell Moonlight do	negative	negative	negative	NN	NN
@Brodylenner i live in manchester er	negative	positive	neutral	NNE	PNE
@jwpage your SMS managed to crash	positive	neutral	negative	PN	NEN
@RedBaff I have popcorn and fruitca	negative	neutral	negative	NN	NEN
@Jessums31 yup its working. they co	positive	neutral	neutral	PNE	NENE
@hazelnutchoc Same as, it's BS I tell	negative	positive	positive	NP	PP
@zachsang when nina picked up my c	negative	neutral	neutral	NNE	NENE
@johan__ thank You! I almost forgot	negative	negative	neutral	NNE	NNE
LAST DAY in St Andrews.. i will spend	negative	neutral	neutral	NNE	NENE
@IislBR My brother shared his microl	negative	neutral	neutral	NNE	NENE
@TheSims3 well us Brits have to wait	negative	positive	negative	NN	PN
Rod stewart's new song	neutral	positive	neutral	NENE	PNE
Roommate is sssnoring :/ And my thr	neutral	negative	neutral	NENE	NNE

Fig. 4. Screenshot of the classification and labelling of tweets.

TABLE III
CONFUSION MATRIX FOR VADER LEXICON

	VADER Lexicon		
	Positive *(Predicted)*	**Neutral *(Predicted)***	**Negative *(Predicted)***
Positive *(Actual)*	22	3	7
Neutral *(Actual)*	4	74	33
Negative *(Actual)*	7	10	140

TABLE IV
CONFUSION MATRIX FOR TEXTBLOB LEXICON

	TextBlob Lexicon		
	Positive *(Predicted)*	**Neutral *(Predicted)***	**Negative *(Predicted)***
Positive *(Actual)*	23	6	3
Neutral *(Actual)*	18	86	7
Negative *(Actual)*	21	27	109

TABLE V
THE PERFORMANCE MEASURE RESULTS OF SENTIMENT ANALYSIS USING TEXTBLOB AND VADER LEXICON

Lexicon	**Polarity**	**Precision**	**Recall**	**F-Measure**	**Accuracy**
VADER	Positive	0.67	0.69	0.68	0.79
	Neutral	0.85	0.67	0.75	
	Negative	0.78	0.89	0.83	
TextBlob	Positive	0.37	0.72	0.49	0.73
	Neutral	0.72	0.77	0.75	
	Negative	0.92	0.69	0.79	

Based on Table V, VADER has a higher accuracy of 0.79 as compared to TextBlob which has an accuracy of 0.73. VADER has the highest precision in classifying neutral tweets with a score of 0.85, the highest recall in negative tweets with a score of 0.89 and the highest F-measure score also for negatives tweets with a score of 0.83. TextBlob on the other hand has the highest precision in classifying negative tweets, the highest recall of 0.77 in neutral tweets and the highest F-measure score of 0.79 in classifying negative tweets. For both lexicons, the scores for the precision, recall and F-measure for the positive polarity is often the lowest with a significant low seen in the precision and F-measure score of 0.37 and 0.49 respectively for TextBlob. From the confusion matrix, it can be seen that although the tweets were selected randomly, the data is imbalanced whereby most tweets have negative polarity. This has caused the positive polarity to achieved low scores.

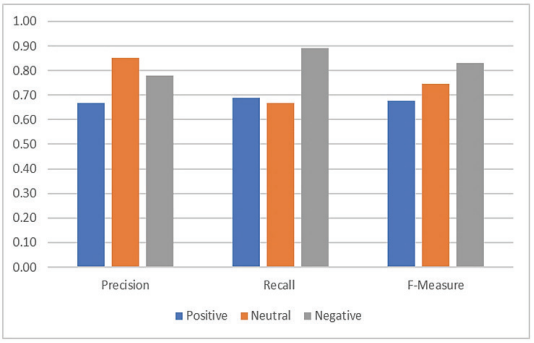

Fig. 5. Comparison of precision, recall and F-measure scores for all polarities using VADER.

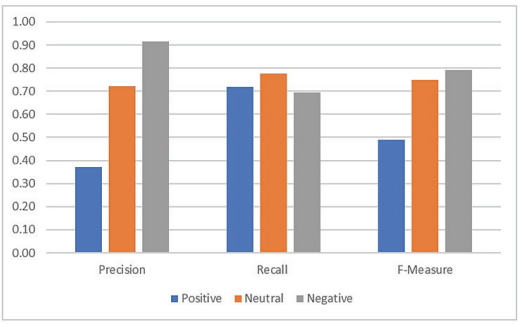

Fig. 6. Comparison of precision, recall and F-measure scores for all polarities using TextBlob.

Comparing both lexicons, VADER achieved a higher score in precision and F-measure for the positive polarity as compared to TextBlob. The poor precision score for TextBlob shows that the lexicon wrongly classifies most of the tweets selected as positive which impacted the F-measure score. For the neutral polarity, VADER has better precision, TextBlob has better recall while both achieved the same score for F-measure. The negative polarity saw VADER achieved a better score in recall and F-measure while TextBlob achieved higher precision score. Fig. 5 presents a bar chart for the precision measure scores for VADER while Fig. 6 presents the same information for TextBlob. Overall, VADER fared better in most polarity and performance metrics as well as achieved a higher accuracy score as compared to TextBlob.

V. Conclusion

The amount of text produced in social media each day allows for various analysis to be performed especially in understanding human behaviour. One of the analysis that can be performed is sentiment analysis. Even though sentiment analysis has been researched for many years, there are still several difficulties in performing it such as in handling internet slangs, abbreviations, and emoticons which is common in social media. In this paper, two lexicons which are VADER and TextBlob were used to compare their efficiency in performing sentiment analysis on Twitter posts. From the study, it is found that both lexicons have an acceptable accuracy rate of 79% for VADER and 73% for TextBlob.

Based on the results of this study, it can be concluded that VADER performs better than TextBlob in classifying tweets. Lexicon-based sentiment analysis is rule-based. Therefore, the performance of the sentiment analysis relies heavily on the quality of the lexicon's dictionary. If a term, for example, "bz" does not exist in the lexicon, it will be ignored hence affecting the polarity score and the sentiment classification of the text. VADER consists of two lexicons which were developed with social media in mind and includes the values of intensity of colloquial and internet slangs. Therefore, more words in the tweets can be correctly analysed by the lexicon as compared to TextBlob hence explaining the result presented in the previous section. It is hoped that this study would aid researches in choosing a lexicon that would perform best in analysing sentiment in social media texts.

Acknowledgment

The authors would like to thank Wan Zaiharatulhasra binti Hamid, Siti Nuur Aisyah bt Dzulkafly and Nurain Nageena Bt Azlizam for evaluating the works produced in this research.

The authors would like to thank the Research Group Optimization Modelling Analytic and Simulation (OPTIMAS) Center For Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Univeristi Teknikal Malaysia Melaka (UTeM)

References

[1] Dream Cyber Infoway PVT LTD. (2019, June 26). Microblogging as a Social Media Initiative towards Successful Business. Retrieved from https://thriveglobal.com/stories/microblogging-as-a-social-media-initiative-towards-successful-business/.

[2] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.

[3] Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.

[4] Padmaja, S., Fatima, S. S., & Bandu, S. (2014). Evaluating sentiment analysis methods and identifying scope of negation in newspaper articles. Int. J. Adv. Res. Artif. Intell, 3(11).

[5] Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools, and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).

[6] Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

[7] Loria, S. (2018). TextBlob Documentation. Release 0.15.

[8] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In Fourth international AAAI conference on weblogs and social media.

[9] Musto, C., Semeraro, G., & Polignano, M. (2014, December). A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts. In DART@ AI* IA (pp. 59-68).

[10] Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science, 44(4), 491-511.

[11] Bonta, V., & Janardhan, N. K. N. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal of Computer Science and Technology, 8(S2), 1-6.

[12] Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), 163-173.

[13] Rao, Y., Lei, J., Wenyin, L., Li, Q., & Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. World Wide Web, 17(4), 723-742.

[14] Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

[15] Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

[16] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

[17] Agerri, R., & García-Serrano, A. (2010, May). Q-WordNet: Extracting Polarity from WordNet Senses. In LREC.
[18] Strapparava, C., & Valitutti, A. (2004, May). Wordnet affect: an affective extension of wordnet. In Lrec (Vol. 4, No. 1083-1086, p. 40).
[19] Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.
[20] Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 241-249). Association for Computational Linguistics.
[21] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. Social science computer review, 29(4), 402-418.
[22] Coppersmith, G., Harman, C., & Dredze, M. (2014, May). Measuring post traumatic stress disorder in Twitter. In Eighth international AAAI conference on weblogs and social media.
[23] Abbar, S., Mejova, Y., & Weber, I. (2015, April). You tweet what you eat: Studying food consumption through Twitter. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3197-3206).
[24] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).
[25] De Smedt, T. & Daelemans, W. (2012). Pattern for Python. Journal of Machine Learning Research, 13: 2031–2035.

Wan Nur Syahirah Wan Min is currently a Junior Data Scientist working on how to use data for valuable business solutions. She received her bachelor's degree in Computer Science (Artificial Intelligence) from University Teknikal Malaysia Melaka, Malaysia. She has a keen interest in building impactful AI solutions through big data analytics.

Nur Zareen Zulkarnain is currently a senior lecturer at Universiti Teknikal Malaysia Melaka, Malaysia. She received her PhD in Computer Science (Natural Language Processing) from the University of Salford, UK. Her research interest includes sentiment analysis, ontology, informatics, and data analytics.