Sentiment analysis on Twitter using machine learning techniques

Anonymous

1. Introduction

With the assistance of universally positioned social network services, it has become possible to generate and share various types of information. As more and more people are sharing their opinions or emotions through social media, a vast amount of data has been accumulated. It becomes an important era to find and utilise information in Big Data. There is a growing need for technologies such as data mining that extract insights from such a large amount of meaninglessly accumulated data to derive meaningful data. Text-based input becomes a common channel for users to share their opinions/emotions through social media [3]. Hence, sentiment analysis using emotions expressed from text-formatted data has been actively studied.

This paper explores text-based emotion prediction using supervised machine learning. The goal is to classify and detect the degree of positivity in 3 sentiments (negative, neutral, and positive) in the texts from Twitter users. For the data pre-processing part, techniques tokenizer and lemmatisation were used. Also, I carried out a process, which changes all emojis found in the text to words that each emoji means, referring to Go, A. (2009). Features were selected by using TF-IDF and BoW methods and hyperparameters were tuned by using grid-search. Supervised machine learning classification algorithms Multinomial Naïve Bayes, Support Vector Machine, Extreme Gradient Boosting, and Decision Tree were performed. Detailed performance of the best model such as accuracy and confusion matrix were also discussed.

2. Related Work

As Big Data develops, studies analysing and predicting sentiment has been increased. In the past, feedback was given to both consumers and companies, limited to formal reviews written by professionals [4]. However, with the development of social media, it has become possible for the general public to freely share their thoughts and feelings. Therefore, a huge amount of emotional data has naturally begun to accumulate, and attempts are made to use it to analyse emotions and develop them. Previous studies have been conducted on representative text-based emotion recognition in three major directions.

The first is to use a morphological analyser. A morpheme means the smallest unit of words that can be separated from the speech flow. Research has studied emotions through the morphological analyser provided by NLTK. It analyses through machine learning with the process of tokenizing the data, and measures accuracy as a division function of phoneme and word units. Research is also underway to proceed with normalization work through recognition of the use pattern or selection rate of words representing extracted emotions.

The second is a method that utilizes machine learning. Previous research has published studies that use machine learning to find utilization rate and patterns of emotional expression words and to analyse emotions in texts. The emotional category has been studied in combination with previous research models, Naïve Bayes and Thayer models. Thayer proposed a numeric-represented dimensional approach that describes the emotional state and classified 12 emotions based on the "Arousal energy degree (High/Low) indicating valence and activity on a scale of positivity and negativity" [5].

Lastly, there is a method of producing and utilizing an emotional-word dictionary. It is mainly constructed with emotional words, emotional classification, and emotional scores. In the past, words with positive meanings were given +1 points, and words with negative meanings were given -1 points. However, this

method had the disadvantage of not being able to reflect the intensity of emotion. Therefore, it has been upgraded using the TF-IDF method, which calculates the frequency of words and calculates the score. By calculating the emotion score in this way, the intensity of emotions can be easily expressed, and a more sophisticated analysis is possible.

Our goal is to distinguish the extent of positivity of emotions in Twitter text and to predict sentiments in phrases that were later written by users.

3. Method

3.1 Feature and Rationale

Emoji

The emojis contained in the text use symbols to represent them directly in the text, so this serves as an important indicator for analysing sentiments. The work of replacing it with an appropriate word according to the meaning of the emoji proceeded. The table shows the example of converting from emojis to words.

Emoji	Word	Emoji	Word
:)	happy	:-s	confuse
:v	funny	:'(	sad
:-o	surprise	:-@	angry
<:o)	party	):	sad

Figure 1- emoji to word

Tokenizer

The task of dividing a given text into units(‘token’) is called tokenization. In ‘Word Tokenization’ used in this paper, tokenization based on whitespace was first performed, then tokenization with only words except for punctuation by recognizing the case of more than one letter or number was done.

Stopwords

Meaningless words include words that appear frequently but are not helpful in the analysis. Here, URLs, HTML tags, digits, hashtags, and mentions are deprecated.

Part-of-speech Tagging

CC (Coordinating junction)	NNS (Noun, plural)	TO (to)
CD (Cardinal number)	NNP (Proper noun, singular)	UH (Interjection)
DT (Determiner)	NNPS (Proper noun, plural)	VB (Verb, base form)
EX (Existential there)	PDT (Predeterminer)	VBD (Verb, past tense)
FW (Foreign word)	POS (Possessive ending)	VBG (Verb, gerund or present participle)
IN (Preposition/subordinating conjunction)	PRP (Personal pronoun)	VBN (Verb, past participle)
JJ (Adjective)	PRPS (Possessive pronoun)	VBP (Verb, non-3rd person singular present)
JJR (Adjective, comparative)	RB (Adverb)	VBZ (Verb, 3rd person singular present)
JJS (Adjective, superlative)	RBR (Adverb, comparative)	WDT (Wh-determiner)
LS (List item marker)	RBS (Adverb, superlative)	WP (Wh-pronoun)
MD (Modal)	RP (Particle)	WPS (Possessive wh-pronoun)
NN (Noun, singular or mass)	SYM (Symbol)	WRB (Wh-adverb)

Figure 2- Speech Codes

In this paper, Part-of-speech Tagging (POS tag) was used to find the part of speech corresponding to a word in the tokenized data. After that, only the parts of speech that can represent the state of emotion were selected.

Unnecessary words were removed by leaving only the words starting with V corresponding to the verb, J corresponding to the adjective, N corresponding to the noun, and R corresponding to the adverb.

Lemmatization

Among the normalization techniques, it refers to a process of searching for a headword (Lemma) from words. Here, the headword (Lemma) means a basic dictionary-type word. It searches the root words that can be generalized and expressed, even if they have different shapes, and uses them to reduce the complexity of the word.

Example	Lemma
am, are, is	be
Studying, studies, study	study
Changed, changer, changing	change
Trouble, troubling, troubled	trouble

Figure 3- example

3.2 Document Term Frequency

Bag of Words

The BoW is a frequency-based word expression method that does not consider the order in which words appear. In this paper, BoW with the CountVectorizer class was created. It is the work of counting the number of appearances by the unit in text for characters with a length of 2 or more and converting it into a numerical vector.

Term Frequency – Inverse Document Frequency

TF-IDF is mainly a task to find the similarity of documents. This means TF multiplied by IDF, where TF means the frequency at which a word is mentioned in a document, and IDF means the frequency at which the word is mentioned in the entire document. In this paper, text was vectorized by adjusting the value to the existing expression using TfidfVectorizer. By tuning the parameters (min_idf, analyzer, sublinear_tf, ngram_range, max_features) in the TfidfVectorizer, accuracy can be increased.

3.3 Classifier – Machine learning Models

- Decision Tree (DT)

Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

- Multinomial Naïve Bayes (NB)

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Support vector machine (SVM)

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, and then the data are transformed in such a way that the separator could be drawn as a hyperplane. SVC was used to proceed with multiple classifications in the process of SVM classification. In terms of the kernel, I used 'RBF' to predict multi-labels.

- Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGB provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.

3.4 Evaluation Method

Confusion Matrix

	Negative (Predict)	Neutral (Predict)	Positive (Predict)
Negative (True)	a	b	c
Neutral (True)	d	e	f
Positive (True)	g	h	i

Figure 4- confusion matrix

The confusion matrix is a table for comparing the predicted value and the actual value to measure the prediction performance through the training process.

Accuracy: $\frac{a+e+i}{a+b+c+d+l+f+g+h+i}$

Weight for f1 score:

$$\frac{1}{\Sigma_{l\in L}|\hat{y}_l|}\sum_{l\in L}|\hat{y}_l|P(y_l,\hat{y}_l)\frac{1}{\Sigma_{l\in L}|\hat{y}_l|}\sum_{l\in L}|\hat{y}_l|R(y_l,\hat{y}_l)$$

4. Result & Discussion

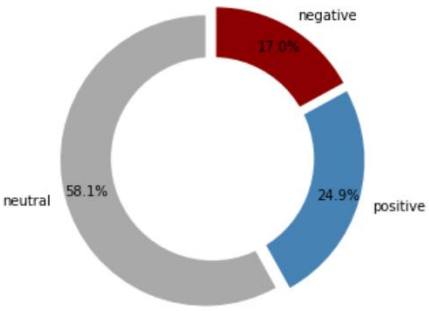

Figure 5- pie chart

Confirming to the pie chart, about 58% of the majority or more were neutral, followed by positives (about 25%), and the smallest percentage was negative (17%). The number or ratio of positive or negative on the train set was somewhat small compared to neutral. Through this, it could anticipate that the results of the modelling could be skewed toward the neutral and could affect accuracy.

Machine Learning Model	Train Accuracy		Test Accuracy	
	BoW	TF-IDF	BoW	TF-IDF
DT	0.6109	0.5806	0.4018	0.3465
NB	0.8349	0.8612	0.5826	0.5199
SVC	0.8562	0.6547	0.5625	0.5236
XGB	0.7909	0.8613	0.5761	0.5232

DT: Decision Tree; NB: Naïve Bayes; SVC: Support vector Classifier; XGB: Extreme Gradient Boosting

Figure 6- Result of modelling with machine learning techniques

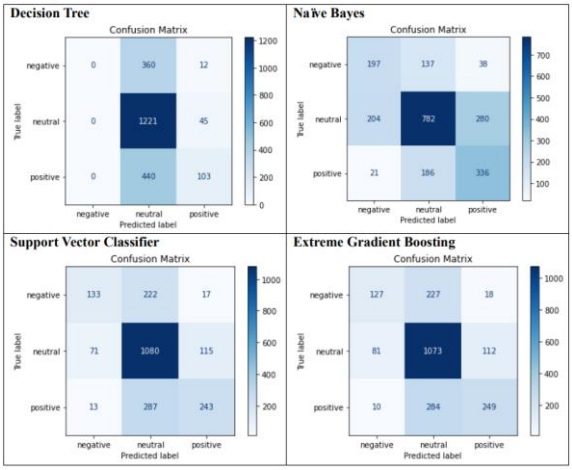

Figure 7- Result using BoW pre-processing

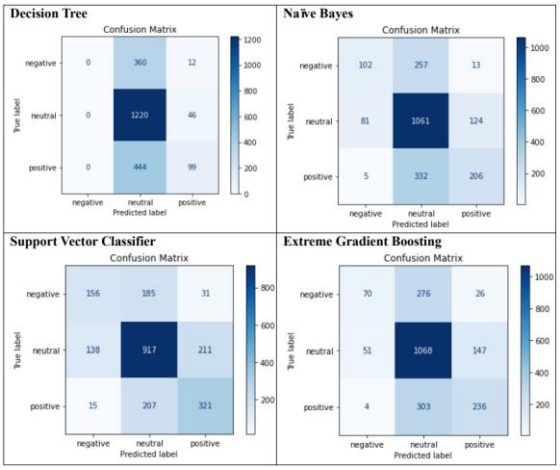

Figure 8- Result using TF-IDF pre-processing

The accuracy of the training dataset of DT was 59.6% on average, and the results performed with the test dataset averaged 37.4%. The result was about 6% higher in the set pre-processed with BoW than the set using TF-IDF. The reason for the low accuracy of DT can be confirmed through the confusion matrix. Since the training data is mostly distributed in neutral, the probability of predicting positive/negative is reduced, and at the same time, in the case of DT, the result of the voting is almost classified as neutral.

The accuracy of NB's training dataset was 84.8% on average, and the average of the results done with the test dataset was 55.1%. The result was about 6% higher in the set pre-processed with BoW than the set using TF-IDF. As a result of performing the entire machine learning model, the result of pre-processing with BoW and modelling with NB was the highest in terms of accuracy, which was about 58.7%. It can be confirmed from the confusion matrix, which the majority of labels are still classified as neutral, but it showed evenly distributed in three labels compared to other models.

The accuracy of the SVC train dataset averaged 75.5%, and the results performed with the test dataset averaged 54.3%. The result was about 3% higher than the set using TF_IDF with one set pre-processed with BoW. In the confusion matrix, I can verify that the SVC model performed the most evenly classified compared to other models pre-processed with TF-IDF.

The accuracy of the XGB's train dataset was 82.6% on average, and the average of the results done with the test dataset was 55%. In the set pre-processed with BoW, the result was about 5% higher than that of the set using TF-IDF.

5. Conclusions

In this paper, emotions in sentences using machine learning models such as Naïve Bayes through texts on Twitter were analysed and predicted sentiments of social media users. The models analysed words in the text and made predictions in three emotional states: positive, negative, and neutral. It was found that the Naïve Bayes model offers the highest accuracy. It showed an accuracy of up to 58.7% and is expected to further contribute to the development of the system in the future. It will be useful for improving the quality of life through various markets in the near future, such as recommender systems, preference analysis, and box-office forecasting.

To get improved accuracy, it is necessary to tune the hyperparameters of machine learning models, but there was a time constraint that I could not try enough attempts. For better results, I recommend pre-processing with a well-constructed emotion dictionary, then going through the modelling or analysing the unbalanced train dataset in consideration of under-sampling/up-sampling.

6. References

[1]Mukherjee, A., Venkataraman, V., Liu, B. & Glance, N. What Yelp fake review filter might be doing? 7th International AAAI Conference on Weblogs and Social Media, 2013.

[2]Rayana, S. & Akoglu, L. Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. 985-994.

[3] Go, A. (2009). Sentiment Classification using Distant Supervision. CS224N project report, Stanford.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 79–86. Association for Computational Linguistics.

[5] Thayer RE (1989) The biopsychology of mood and arousal, Oxford University Press, New York, New York; Oxford, England.