COMP90049 Knowledge Technologies Project 2 Report

Fengke Song

1. Introduction

Nowadays, as the development of data mining, the concept of machine learning is proposed to enable machines to learn features of a series of data and then make predictions on their own. In particular, this project is required to adopt supervised learning methods to predict which emoji is possible used in a tweet. This report will introduce two supervised learning methods and analyse their performance based on the given data set.

1.1 Problem Review

Generally, the project provides a deal of training samples which are used to train different classifiers. Namely, each sample consists of an ID number, a concrete tweet text and its corresponding emoji symbol. There are totally ten unique classes of emoji such as Clay, Cry and Disappoint. Moreover, the key task is to train classifiers which may use various classification methods. After building the specific classifiers, it is necessary to test and verify their performance through testing quantities of testing samples and evaluating the result by accuracy, confusion matrix, precision, recall and F1 score.

1.2 Data Set Analysis

The raw data set can be divided into four main types which are training data, testing data, feature words and high frequency words respectively. Both training and testing data are collected from Twitter through its API^2 (https://developer.twitter.com/, 2018), which means there may exist several mistakes:

(1) There are various misspelling issues in the training data. This problem may influence the selection of features because whether one word could become a feature probably depends on its appearing frequency.

(2) Many tweets have specific links which will jump to various news or articles. Nevertheless, the emojis these tweets use may rely on the content of these links which cannot be analysed by methods.

(3) There are ten different kinds of emoji selected as the classes. However, it is not easy for humans to exactly classify a sentiment text among so many

classes, let alone for machines.

(4) As Table 1 shows, the number of samples belonged to each class is heterogeneous, which might influence the precision of predicted results for different classes.

Clap	Cry	Disappoin t	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
1265	1587	461	1334	433	1322	1496	1222	1420	1624

Table 1. The number of samples belonged to each class

As a result, all these mistakes may influence the performance of supervised learning methods. Besides, the number of feature words also affects the accuracy of every classifier. There are 100 features given by the project which are determined through mutual information and chi-square methods. However, to train a high performing classifier, the number is far from enough. Besides, the high frequency words include many useless words which should be stopped from adoption such as ‘an’, ‘the’ and so on. Consequently, this report will select more features by TF-IDF methods to evaluate the effect of different features.

2. Relevant Literature

Undoubtedly, there has existed several researches about Twitter’s sentiment analysis. Besides, almost all literatures adopt supervised learning methods to train their classifiers such as Naïve Bayes and Decision Tree (Oscar, Fox, Croucher, Wernick, Keune, & Hooker, 2017). However, those researchers prefer to classify sentiment texts as two main classes: positive and negative. In specific, sentiment like smile, happy and spirit are positive ones, while down, complaint and angry are negative ones (Susanti, Djatna, & Kusuma, 2017). Consequently, if features could be well selected and classifiers could be perfectly trained, they would get quite high accuracy of the results around 80%.

3. Supervised Learning Methods

Supervised learning is mainly applied to generate specific classifiers through training thousands of labelled samples, and new instances could be predicted automatically. This report will describe two supervised learning methods which are Naïve Bayes Classifier and K-Nearest Neighbors Classifier.

3.1 Naïve Bayes Classifier

As is explained in Knowledge Technologies Lecture 4 of Part B (Jeremy, Justin, Karin, &Rao, 2016), Naïve Bayes Classifier is a supervised learning method based on probability. In particular, given a series of n dimensional training attribute vector $X = (x_1, x_2, \ldots, x_n)$ and their k associated classes $C_1, C_2, \ldots, C_k$, the probability of a targeted testing vector X for each class C_i is $P(C_i|X)$. According to the Naïve Bayes Formula $P(C_i|X) = P(X|C_i)P(C_i)/P(X)$, as $P(X)$ is the evidence and has no effect to the result, the probability $P(C_i|X)$ depends on $P(X|C_i)P(C_i)$. That means the class with maximized $P(C_i|X)$ is the most possible classification for the targeted testing vector X.

In addition, it is essential to assume that each attribute is conditionally independent. Therefore, $P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$.

3.2 K-Nearest Neighbors Classifier

K-Nearest Neighbors (KNN) Classifier is a distance based supervised learning method. Similarly, assuming that there are m training samples, and each sample i with n dimensional attribute vector $X_i = (x_1, x_2, \ldots, x_n)$ and their k associated classes $C_i \in (C_1, C_2, \ldots, C_k)$. To determine which class a testing vector $T = (t_1, t_2, \ldots, t_n)$ belongs to, the key idea is calculating the distances between T and each X_i. Afterward, sorting all distances increasing order, and selecting the front K entries. The class that appears among these K entries with the most times is the most possible classification for T.

4. The Selection of Features

Undoubtedly, the selection and the number of features used to train classifiers will influence the result of every method. This report will begin by adopting the 100 features given by the project which are determined by mutual information and chi-square. After that, the report will continue to research the influence of features through using the features which are determined by TF-IDF and increasing the number of features at intervals of 200 (100, 300, 500…). Namely, most of the stop words (https://www.ranks.nl/stopwords, 2018) will be excluded from the features.

5. The Parameter of Classifiers

Actually, the Naïve Bayes classifier has no parameter that would influence its performance. However, the KNN classifier owns two parameters which may influence its performance. The first one is the method used to calculate the distance. This report will choose Euclidean distance as the formula which is $d = \sqrt{(T - X_i)^2}$ (http://www.pbarrett.net, 2005). The second one is the selection of K. The results contributed by different K will be researched in this report.

6. The Performance of Classifiers

This report will adopt accuracy, confusion matrix, precision, recall and f1_score to evaluate the performance of each classifier. The concrete calculating formulas are presented in Knowledge Technologies Lecture 4 of Part B (Jeremy, Justin, Karin, &Rao, 2016).

6.1 The Performance of Naïve Bayes Classifier

1. When adopting given 100 features determined by mutual information and chi-square, the performance of this classifier is:

(1) The accuracy is 27.01%;

(2) The confusion matrix is shown in Table 2;

Actual / Predict	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Clap	649	419	44	433	72	1496	198	1420	423	170
Cry	58	273	29	86	25	43	116	72	66	154
Disappoint	5	1	73	4	1	3	8	1	6	13
Explode	142	133	12	445	49	55	78	74	115	46
FacePalm	10	4	2	2	45	3	4	2	2	1
Hands	27	4	1	9	4	323	7	7	18	15
Neutral	5	51	6	20	7	10	148	22	15	4
Shrug	7	44	20	20	9	7	68	138	36	91
Think	4	7	1	0	1	1	1	2	99	2
Upside	358	642	273	465	220	295	871	715	640	1092

Table 2. The confusion matrix of Naive Bayes Classifier with top10 features

(3) The precision, recall and f1_score are shown in Table 3;

	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Precision	21.49%	29.61%	54.89%	38.73%	60.00%	78.40%	45.68%	31.36%	83.90%	19.60%
Recall	51.30%	17.20%	15.84%	33.36%	10.39%	24.43%	9.89%	11.29%	6.97%	67.24%
F1_Score	0.30	0.22	0.25	0.36	0.18	0.37	0.16	0.17	0.13	0.30

Table 3. The Precison, Recall and F1_Score of Naive Bayes Classifier

2. When adopting 100 features which are determined by TF-IDF method, the result is:

(1) The accuracy is 29.69%;

(2) The confusion matrix is shown in Table 4;

Actual Predict	Clap	Cry	Disappoi nt	Explode	FacePal m	Hands	Neutral	Shrug	Think	Upside
Clap	725	479	78	208	135	425	244	204	321	188
Cry	74	303	37	61	16	37	67	52	43	75
Disappoi nt	3	14	9	3	2	4	8	1	2	10
Explode	8	133	2	391	2	55	11	19	115	12
FacePal m	1	4	0	2	0	2	0	2	2	2
Hands	67	30	5	20	12	554	13	20	50	47
Neutral	47	136	45	112	43	28	300	160	111	185
Shrug	20	43	63	53	15	8	94	153	63	128
Think	65	110	17	70	27	62	97	104	317	117
Upside	255	456	205	415	181	176	663	508	500	860

Table 4. The confusion matrix of Naive Bayes Classifier with 100 TF-IDF features

(3) The precision, recall and f1_score are shown in Table 5;

	Clap	Cry	Disappoin t	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Precision	24.11%	39.61%	16.36%	78.67%	0.00%	67.73%	25.71%	23.91%	32.15%	20.38%
Recall	57.31%	19.09%	1.95%	29.31%	0.00%	41.91%	20.05%	12.52%	22.32%	52.96%
F1_Score	0.34	0.26	0.03	0.43	0.00	0.52	0.23	0.16	0.26	0.29

Table 5. The Precison, Recall and F1_Score of Naive Bayes Classifier

3. When adopting 300 features which are determined by TF-IDF method, the result is:

(1) The accuracy is 37.33%;

(2) The confusion matrix is shown in Table 6;

Actual Predict	Clap	Cry	Disappoi nt	Explode	FacePal m	Hands	Neutral	Shrug	Think	Upside
Clap	760	339	40	161	68	324	152	152	227	141
Cry	110	569	64	111	30	80	118	104	112	122
Disappoi nt	9	23	104	8	4	3	15	17	9	19
Explode	18	42	7	500	16	31	40	39	28	45
FacePal m	17	7	4	2	67	4	7	5	8	3
Hands	59	27	3	64	12	658	11	17	20	39
Neutral	42	91	42	83	49	41	397	144	119	193
Shrug	28	83	31	70	29	8	132	285	124	184
Think	29	41	13	25	18	25	57	58	380	57
Upside	193	365	153	310	140	148	567	401	393	821

Table 6. The confusion matrix of Naive Bayes Classifier with 300 TF-IDF features

(3) The precision, recall and f1_score are shown in Table 7;

	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Precision	32.15%	40.07%	49.29%	65.27%	54.03%	72.31%	33.06%	29.26%	54.05%	23.52%
Recall	60.08%	35.85%	22.56%	37.48%	15.47%	49.77%	26.54%	23.32%	26.76%	50.55%
F1_Score	0.42	0.38	0.31	0.48	0.24	0.59	0.29	0.26	0.36	0.32

Table 7. The Precison, Recall and F1_Score of Naive Bayes Classifier

4. When adopting 500 features which are determined by TF-IDF method, the result is:
(1) The accuracy is 39.97%;
(2) The confusion matrix is shown in Table 8;

Actual / Predict	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Clap	540	107	19	27	21	177	56	60	74	76
Cry	284	793	70	181	47	168	154	174	165	166
Disappoint	10	27	105	13	5	3	19	15	9	27
Explode	56	63	14	569	19	51	60	35	53	63
FacePalm	17	10	2	6	88	4	10	5	8	3
Hands	59	29	4	72	9	700	15	24	36	41
Neutral	49	96	45	91	65	49	495	140	125	205
Shrug	35	88	33	85	37	18	164	352	147	217
Think	40	46	16	33	15	20	57	69	460	57
Upside	168	328	153	257	127	132	466	348	340	760

Table 8. The confusion matrix of Naive Bayes Classifier with 500 TF-IDF features

(3) The precision, recall and f1_score are shown in Table 9;

	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Precision	46.67%	36.01%	45.45%	57.88%	52.69%	70.28%	36.40%	29.93%	56.58%	24.68%
Recall	42.69%	49.97%	22.78%	42.65%	20.32%	52.95%	33.09%	28.81%	32.39%	46.80%
F1_Score	0.45	0.42	0.30	0.50	0.29	0.60	0.35	0.29	0.41	0.32

Table 9. The Precison, Recall and F1_Score of Naive Bayes Classifier

5. When adopting 100 features determined by mutual information and chi-square combined with 500 features which are determined by TF-IDF

method, the result is:

(1) The accuracy is 40.98%;

(2) The confusion matrix is shown in Table 10;

Actual / Predict	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Clap	702	210	22	117	42	267	96	117	140	90
Cry	115	658	64	90	30	75	110	107	95	149
Disappoint	7	28	107	17	6	3	23	15	8	22
Explode	74	97	16	637	36	42	88	47	75	82
FacePalm	17	9	2	7	83	4	11	5	12	8
Hands	71	27	5	22	7	726	9	22	44	35
Neutral	49	116	48	87	63	38	482	148	136	205
Shrug	32	75	35	76	32	16	162	335	136	198
Think	41	42	10	33	16	26	49	72	468	48
Upside	157	325	152	248	118	125	466	354	306	787

Table 10. The confusion matrix of Naive Bayes Classifier with mixture features

(3) The precision, recall and f1_score are shown in Table 11;

	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
Precision	38.94%	44.07%	45.34%	53.35%	52.53%	75.00%	35.13%	30.54%	58.14%	25.91%
Recall	55.49%	41.46%	23.21%	47.75%	19.17%	54.92%	32.22%	27.41%	32.96%	48.46%
F1_Score	0.44	0.43	0.31	0.50	0.28	0.63	0.34	0.29	0.42	0.34

Table 11. The Precison, Recall and F1_Score of Naive Bayes Classifier

6.2 The Evaluation of Naïve Bayes Classifier

As is presented in 5.1, though the accuracy of each condition is not quite high, the Naïve Bayes classifier could predict many testing samples. Importantly, when adopting 100 features determined by mutual information and chi-square combined with 500 features which are determined by TF-IDF method to train the classifier, the accuracy is up to nearly 41%, and over half of the 10 classes own fairly high f1 scores (more than 0.41). Besides, the heterogeneity of the number of the training samples belonged to each class restricts the accuracy of prediction. For example, the number of training samples belonged to 'FacePalm' is 433, its f1 score is just 0.28 which is quite low.

In terms of how features influence the results, Table 12 and Table 13 illustrate the changes.

	100 features (mutual information)	100 features (TF-IDF)	300 features (TF-IDF)	500 features (TF-IDF)	100 (mutual information) and 500 (TF-IDF)
Accuracy	27.01%	29.69%	37.33%	39.97%	40.98%

Table 12. The accuracy of different features

F1_Score	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
100 (mutual information)	0.30	0.22	0.25	0.36	0.18	0.37	0.16	0.17	0.13	0.30
100 (TF-IDF)	0.34	0.26	0.03	0.43	0.00	0.52	0.23	0.16	0.26	0.29
300 (TF-IDF)	0.42	0.38	0.31	0.48	0.24	0.59	0.29	0.26	0.36	0.32
500 (TF-IDF)	0.45	0.42	0.30	0.49	0.29	0.60	0.35	0.29	0.41	0.32
100 (MI) and 500 (TF-IDF)	0.44	0.43	0.31	0.50	0.28	0.63	0.34	0.29	0.42	0.34

Table 13. The F1_Score of different features

In Table 12, comparing the accuracy whose features are determined by mutual information and chi-square with the one whose features are determined by TF-IDF, it is clear that different feature selection methods lead to different accuracy. Moreover, while using the same method TF-IDF to select features, the accuracy still increases in pace with the steadily increasement of the number of features. Table 13 gives various details of f1 scores for these 10 classes. Although there exist several fluctuations while changing or increasing the features, the trend of the results still gets better.

6.3 The Performance of K-Nearest Neighbors Classifier

In fact, in terms of the influence of different features, KNN classifier has the same trend as Naïve Bayes classifier. Therefore, this report will focus on comprising of the performance between these two classifiers.

As Table 14 demonstrates, the accuracy of KNN (K = 100) classifier is similar with that of Naïve Bayes classifier. Besides, the performance of KNN classifier is also restricted by the number of training samples. For instance, in Table 15, KNN classifier also owns poor performance on class ‘Disappoint’ whose sample number is only 461.

	100 features (mutual information) NB	100 features (mutual information) KNN	100 features (mutual information) and 500 features (TF-IDF) NB	100 features (mutual information) and 500 features (TF-IDF) KNN
Accuracy	27.01%	27.63%	40.98%	40.83%

Table 14. The accuracy comparison between NB and KNN

F1_Score	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
100 (mutual information) NB	0.30	0.22	0.25	0.36	0.18	0.37	0.16	0.17	0.13	0.30
100 (mutual information) KNN	0.23	0.30	0.27	0.38	0.19	0.39	0.26	0.15	0.24	0.30
100 (MI) and 500 (TF-IDF) NB	0.46	0.43	0.31	0.50	0.28	0.63	0.34	0.29	0.42	0.34
100 (MI) and 500 (TF-IDF) KNN	0.44	0.43	0.26	0.53	0.32	0.70	0.34	0.24	0.37	0.35

Table 15. The F1_Score of different features and classifiers

6.4 The Analysis of K-Nearest Neighbors Classifier

This report will analyse whether changing the value of K would influence the performance of KNN classifier. Specifically, the following analysis will base on the 100 features determined by mutual information and chi-square, and K can be 50, 100 and 200.

	K = 50	K = 100	K = 200
Accuracy	26.70%	27.63%	27.43%

F1_Score	Clap	Cry	Disappoint	Explode	FacePalm	Hands	Neutral	Shrug	Think	Upside
K = 50	0.29	0.24	0.27	0.35	0.19	0.39	0.18	0.16	0.24	0.31
K = 100	0.23	0.30	0.27	0.38	0.19	0.39	0.26	0.15	0.24	0.30
K = 200	0.22	0.30	0.26	0.37	0.19	0.39	0.24	0.20	0.24	0.31

Table 16. The influence of K on KNN classifier

Table 16 illustrates that the influence on the performance of KNN classifier caused by K can be ignored because both the accuracy and f1 score fluctuate slightly.

7. Conclusion

The report has demonstrated that both Naïve Bayes classifier and K-Nearest Neighbors classifier could do well in classification. That means the supervised learning methods are really able to automatically classify sentiment texts and do predictions with quite high accuracy. Besides, the main factors that influence the performance of predictions are the selection of features and the number of features. Furthermore, the number of testing samples and the number of classes also affect much on the results. If providing more samples of each class and further optimizing features, both classifiers will get better results.

References

Euclidean Distance 'raw, normalized, and double-scaled coefficients'. (2005). Retrieved from http://www.pbarrett.net/techpapers/euclid.pdf.

Jeremy N., Justin Z., Karin V., & Rao K. (2016). COMP90049 Knowledge Technologies *Lecture4 of Part B Introduction Classification*, The University of Melbourne.

Oscar, N., Fox, P. A., Croucher, R., Wernick, R., Keune, J., & Hooker, K. (2017). Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. *Journals of Gerontology Series B: Psychological Sciences & Social Sciences*, 72(5), 742-751.

Ranks NL Stop Words. (2018). Retrieved from https://www.ranks.nl/stopwords.

Susanti A. R., Djatna T., & Kusuma W. A. (2017). *Twitter's Sentiment Analysis on Gsm Services using Multinomial Naïve Bayes.* Telkomnika pp. 1354.

Twitter Developer Docs. (2018). Retrieved from https://developer.twitter.com/en/docs/tweets/search/.