**Intro**

**Recite(100)**

The goal of this machine learning task is to correctly predict the sentiment given a twitter text. This is a task to find the most suitable mapping function from feature to label, therefore, the supervised machine learning model would be capable for this task.

This project follows 4 general steps. First process is preprocessing and transform the raw dataset in order to obtain a much cleaner and machine-readable data. Second process is features engineering. By selecting the most relevant features, the noise and dimensionality of models get reduced. Third process is training various models and using the models to prediction the test data. Last process is evaluating and analyzing every model.

The raw training dataset contain 2 attributes, id and text, and 1 label, sentiment.

Features:

The ID feature contains the information about the user numerical ID. Another feature is the text post by users. The text feature contain a single string text which contains both useful information and noise need to be clean.

For the label, the sentiment are given for every training instance. The distribution of the value counts which are showed by bar chart suggests that the distribution of the data is highly imbalanced. There are 12659 neutral instances, 5428 positive instances and 3715 negative instances. The number of neutral instances is much larger than the number of positive instances and negative instances combined. The imbalanced distribution of data could be a potential hinder for the data modeling.

**Method**

In this section, we elaborate several

To begin with, the preprocess of the raw data get perform. There are two subprocess get implemented during the preprocess, one is data cleaning., another one is vectorization.

Data cleaning (could write more)

The goal of data cleaning is to only include the relevant information by reducing the noisy data. There are two features in the raw data set correspond with one label. However, The ID feature are not chosen to conduct the supervised machine learning models, because the numeric account ID itself does not contain sentimental information. The reason of including the text feature in the raw dataset to perform the supervised learning is that only the texts post by users are able to be analyzed.

To further clean the text attribute, three kinds of texts get removed or transformed. There are words which are irrelevant to sentiment, the same words with different form and the different words with the same lexeme.

Firstly, the irrelevant text gets removed including Special words such as @username, #tag, URL., punctuations, numbers and stop word such as is, or, and. For example, special words like #china should contain no emotional information, because only the comments about china contain sentiment, but the tag itself do not.

Secondly, the same words with different form get transform into one base form by expanding the contraction, replacing the accented characters, lowering all cases and converting the number into word form. For instance, the abbreviation, “I’m”, gets expanded into two sperate words. “I” and “am”. The reason of transforming into a base form is that those words are the same ‘I’m’ and ‘I’, ‘am’ are the words with different form, they are the same grammatically. So, they should be classified into the same words instead of classifying they into sperate words.

Thirdly, the different words have the same lexeme get replace by the basic lexeme form. By do the lemmatization, words like ‘is’, ‘are’ get reduce to the lexeme, ‘be’. So that the words with the same lexeme could be classified as the identical words.

transformation

In addition to the preprocessing, the text features need to transform into the machine-friendly form. There are two common transformation methods, bag of words and term frequency – inverse document frequency (TF-IDF).

Therefore, the TFIDF vectorization is implemented.

And two-gram model are used for the vectorizer to make the model to identify the different between one word and the negation of that word. For example, ‘like’ and not ‘like’ could potentially have opposite sentiment. The one-gram model could not identify them correctly, since every word are be counted individually, but the 2-gram model receive more complete message of the text which be potentially beneficial for sentiment identification.

The data preprocess reduces the dimensionality of transformed feature significantly from 204657 to 127482.

Feature engineering (150)

**Modelling:**

Train test split

Starting with splitting the training set into two subsets by keeping the size of one subset same as the testing dataset. One of the split data is for the training and another section which have 6099 instances of data is for the validation. The reason of keeping 6099 instances in validation set is to better simulate the testing data. Random holdout method for data splitting is chosen for this task despite the k-fold cross validation would be the most preferable method. The choose the suboptimal method is mainly caused by the insufficient computation power when dueling with large dataset.

There are three types of model being used to model the data and compare the performance.

First of all, the baseline model 0R get built. 0R model constantly predict the most frequent labels. And base model works as a baseline to compare with all other models to measure the performance.

Secondly, several base models are used to compare the performance and built for following ensemble stacking model. They are decision tree, logistic regression model, support vector machine. Decision tree is a simple machine learning model for categorical features. The advantages of decision tree are simple implementation, high interpretability and no distribution are assumed. However, the major drawback is overfitting. To be more specific, decision tree has high variance and low bias, which lead to poor performance for new data input. Logistic regression uses the probability of success and failure of event to build the classification model(geekforgeek). The major benefits of implementing logistic model are the same as the decision tree, but the disadvantage is that the model without transforming the data can only detect the linear relationship between features and labels. Support vector machine (SVM) are benefit by effective in high dimensional spaces and memory efficient (Dhiraj K). Whereas, the advantages of implementing SVM are highly affect by noise.

To tuning the parameter for every base model, the result from the default setting and best result from the SearchCV algorithm are compared to obtain the combination of hyperparameters with better score. There are two SearchCV algorithm, RandomizedSearchCV and GridSearchCV. Considering the feasibility of this project, randomized search method is better suit for this situation instead of grid search, because the dataset is considerably large. Although the grid search guarantees the parameters with best score, the exhaustive grid search can consume a huge amount of computation power and time which make the project potentially infeasible. Therefore, the randomized search method with 10% samples of the search space of hyperparameter are used to approximate the best parameter. By doing the 10% random sampling, the computation power and time is saved, and the potential overfitting problem is avoided by having sub optimal solution. In addition to scoring each parameter combination for RandomizedSearchCV function, ‘f1\_weighted’ are used to measure the performance. There are two major advantages of setting f1\_weighted for this imbalanced dataset. One advantage is that weighted score takes the numbers of each labels into consideration. The more frequent label will take more weight and the impact of the uneven data get reduced. Another advantage is that f1 score encourages the model to balance the precision and recall. F1 score measures both correctness of predicting positive cases and the ability to detect the true positive (lecture sides 6).

() (formula for precision and recall)

As the formula shows above, the f1 score is the combination of the precision and recall.

The precision gives the ratio of true positive out of all predicted positive, and recall give the ratio of true positive out of all correctly predicted labels. By combing precision and recall, f1 score measures both correctness of predicting positive cases and the ability to detect the true positive (lecture sides 6). Whereas other scoring method like accuracy can be strongly biased by the majority label, which can result in predicting one class all the time.

Thirdly, more complex model using ensemble learning methods get built. Both of bagging and stacking models combining several weak leaners to obtain a stronger and more sophisticated model. To be specific, the data get modeled by the random forest model which is a typical bagging model and stacking models with several different meta-classifiers.

Random forest combines result from several decision trees to produce the prediction. The advantages of random forest algorithm comparing with decision tree are the chance of overfitting get reduced.

For stacking model, the combination of the base models from second step are used as the basis for stacking. Three metaclassifiers, logistic regression and support vector machine are implemented for comparison. The benefit of using stacking is that the ensembled model are combining several well-performing models together and usually have better practical result(). However, for the ensemble algorithm, both time and space complexity become enormous which is resulted by training multiple base learner and then using metaclassifier to train the final model.(). In addition to drawback of both random forest and stacking models, models are hard to interpret and impossible to be fully understandable, the bagging models and stacking models perform reasonably good in this task, and the goal of supervised machine learning is to get more correct prediction rather than interpret the logic behind the relations(lec slide).

**Result (400)**

The final model get chosen is the stacking model with logistic metaclassifier. The model have 0.71 validation accuracy, 0.725 weighted f1 score (0.724 precision and 0.73 recall).

(interpret the metrics)

**Discussion(600)**

**Impact of the imbalance dataset and effort to tackle the imbalance(163/200)**

Observing the raw training dataset, the imbalance distribution of the data could be found. Three sentiment group have huge difference in terms of the total number where neutral labels have 12659 cases, positive labels and negative labels have 5428 and 3715 instances respectively. The neutral sentiment instances are more than all other two instances combined. This uneven number of instances from different group could lead to Low recalls and precision for the minority groups. Because of the imbalanced data structure, models are encouraged to predicting the neutral labels. By predicting the majorities and ignore minorities (0R model), the model accuracy could reach reasonable accuracy, 0.5. Because models will predict less minorities, the recall or precision decreases significantly. This For example, 0R model which is the extreme case of predicting majority label have low precision recall and f1, 0.346, 0.588, 0.435 respectively. The solution comes up by our search group are implementing balanced class weight parameter for modelling and using weighted f1 as scoring.

(comparison before and after balance class weight)

**High FN and FP problem (400)**

One problem shows by the confusion matrices is that most of modelling method have both high false negative and false positive for neutral labels. Models can identify the difference between negative and positive class., but models perform poorly for identifying difference between neutral and positive class and difference between neutral and negative class.

There are two potential hypothesis that comes by our groups, one hypothesis relates to the text input, another hypothesis relates to the parameter tuning. The first hypothesis of having both high FN and FP rate is that the neutral class shares many identical words with positive and negative classes. When model tries to predict new labels, those shared words confuse the model, and predict with the wrong label. The second hypothesis is that the sub optimal approximation produced by the RandomizedSearchCV is not close enough to the true optimal hyperparameter. To reduce the time complexity, only most relevant portion of parameter identified by our research group. There is chance that some other important hyperparameter not get considered. In addition to second hypothesis, another problem might be caused by the RandomizedSearchCV is that search algorithm only randomly searches for a small proportion of the all hyperparameter sample space. This insufficient sample size problem also could make that the incorrect approximation of the optimal hyperparameters.

(150)

Some potential solutions are that increases the number of instances for two minority classes, apply resampling methods and change the class weight. one of the best solution for solving high FN and FP is that increase the instances of positive and negative sentiment cases. By increasing the sample size of two minority classes, there are higher chance that new deterministic features are learned by models which could improve the FN and FP.

Second, resampling method,

Thirdly, changes of class weight could also potentially benefit the prediction. For the simplicity, most of models are using ‘balanced’ parameter which takes the inverse proportion of the frequency, but other weighted strategies could also be tried.

**Conclusion(300)**

**Bibliography**

[**https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec#:~:text=Some%20advantages%20of%20F1%2Dscore,metric%20across%20positive%2Fnegative%20samples**](https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec#:~:text=Some%20advantages%20of%20F1%2Dscore,metric%20across%20positive%2Fnegative%20samples)**.**

**https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/**

[**https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107**](https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107)

**http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html**

**https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/#:~:text=The%20benefit%20of%20stacking%20is,single%20model%20in%20the%20ensemble.**