

Categorization of Users News Interest Based on Tweets

V. Akshay (2015503005)

Department of Computer Technology
Madras Institute of Technology
Anna University
chennai - 44, India
akshayred@gmail.com

V. Gowtham (2015503517)

Department of Computer Technology
Madras Institute of Technology
Anna University
chennai - 44, India
gowthamv441@gmail.com

Praveen Siva (2015503540)

Department of Computer Technology
Madras Institute of Technology
Anna University
Chennai - 44, India
praveen17siva@gmail.com

Abstract—This Project mainly aims at the categorization of the user's interest based on the current news on which the user is tweeted. This enables to provide the user interested news which eventually increase the user enhanced environment.

Keywords—NLP, linear svm, Multinomial NB, Logistic Regression, Random Forest Classifier.

I. INTRODUCTION

There are heaps of uses of content characterization in the business world. For instance, news stories are ordinarily composed by subjects; substance or items are regularly labeled by classes; clients can be grouped into associates dependent on how they discuss an item or brand online. However, by far most of content arrangement articles and instructional exercises on the web are paired content characterization, for example, email spam sifting (spam versus ham), opinion examination (positive versus negative). As a rule, our genuine issue are considerably more confounded than that. Hence, this is the thing that we will do is Classifying debacle enthusiasm into 7 predefined classes. Therefore, form categorizing user interest based on his tweets using his tweet w can provide the user interested news over the natural disaster event so that we can provide the user enhanced environment.

II. RELATED WORK

A. Natural Language processing

The field of concentrate that centers around the collaborations between human dialect and PCs is called Natural Language Processing, or NLP for short. It sits at the crossing point of software engineering, computerized reasoning, and computational phonetics. NLP is a path for PCs to examine, comprehend, and get significance from human dialect in a keen and helpful way. By using NLP, engineers can arrange and structure learning to perform undertakings, for example, programmed outline, interpretation, named substance acknowledgment, relationship extraction, feeling investigation, discourse acknowledgment, and point division.

B. Linear SVM

In machine learning, bolster vector machines (SVMs, additionally bolster vector networks) are regulated learning models with related learning calculations that dissect information utilized for order and relapse investigation. Given an arrangement of preparing precedents, each set apart as having a place with either of two classifications, a SVM preparing calculation fabricates a model that allocates new models to one classification or the other, making it a non-probabilistic parallel direct classifier (in spite of the fact that

strategies, for example, Platt scaling exist to utilize SVM in a probabilistic characterization setting). A SVM demonstrate is a portrayal of the models as focuses in space, mapped so the precedents of the different classifications are separated by an unmistakable hole that is as wide as could reasonably be expected. New models are then mapped into that equivalent space and anticipated to have a place with a classification dependent on which side of the hole they fall.

C. Multinomial NB

In machine learning, guileless Bayes classifiers are a group of straightforward "probabilistic classifiers" in view of applying Bayes' hypothesis with solid (innocent) freedom suppositions between the highlights. Guileless Bayes has been considered broadly. It was brought under an alternate name into the content recovery network, remains a well known (benchmark) strategy for content arrangement, the issue of passing judgment on archives as having a place with one class or the other, (for example, spam or authentic, sports or legislative issues, and so on.) with word frequencies as the highlights. With fitting pre-handling, it is focused in this area with further developed techniques including bolster vector machines. It additionally discovers application in programmed therapeutic analysis.

D. Logistic Regression

In insights, the calculated model (or logit display) is a broadly utilized factual model that, in its essential shape, utilizes a strategic capacity to show a parallel ward variable; numerous more mind-boggling expansions exist. In relapse investigation, strategic relapse (or logit relapse) is evaluating the parameters of a calculated model; it is a type of binomial relapse. Scientifically, a twofold strategic model has a needy variable with two conceivable qualities, for example, pass/fall flat, win/lose, alive/dead or solid/wiped out; these are spoken to by a pointer variable, where the two qualities are marked "0" and "1".

E. Random Forest Classifier

Random Forest or arbitrary choice forest are a gathering learning technique for grouping, relapse and different errands, that work by building a large number of choice trees at preparing time and yielding the class that is the method of the classes (order) or mean expectation (relapse) of the individual trees. Random choice timberlands revise for choice trees' propensity for overfitting to their preparation set. The first calculation for arbitrary choice woodlands was made by Tin Kam Ho utilizing the irregular subspace strategy, which, in Ho's plan, is an approach to execute the "stochastic separation" way to deal with arrangement.

III. ALGORITHM INVOLVED

In machine learning, multiclass or multinomial order is the issue of characterizing cases into one of at least three classes. (Characterizing occasions into one of the two classes is called double characterization.) While some order calculations normally allow the utilization of in excess of two classes, others are commonly twofold calculations; these can, in any case, be transformed into multinomial classifiers by an assortment of techniques. Multiclass arrangement ought not be mistaken for multi-name grouping, where different marks are to be anticipated for each occurrence.

A. Abbreviations and Acronyms

Here is the unavoidable abbreviation in the project they are SVM stands for Support Vector Machines, MNB stands for Multinomial Naïve Bayes, NLP stands for the Natural Language Processing.

B. Problem Formulation

The issue is managed content grouping issue, and our objective is to examine which regulated machine learning strategies are most appropriate to settle it. Given another tweets comes in, we need to appoint it to one of 7 classifications. The classifier makes the supposition that each new tweet is relegated to one and just a single classification. This is multi-class content characterization issue.

C. Data Exploration

Before jumping into preparing machine learning models, we should take a gander at a few precedents first and the quantity of grievances in each class. For this task, we require just two columns—"tweet" and "catastrophe label". Input is tweet. Precedent: " I have obsolete data on my credit report that I have recently debated that still can't seem to be expelled this data is all the more than seven years of age and does not meet credit revealing prerequisites". Yield: "Random". Example: off-topic. We will expel missing qualities in "tweet" section, and include a segment encoding the item as a whole number in light of the fact that straight out factors are regularly preferable spoken to by numbers over strings. We likewise make a few lexicons for sometime later.

D. Imbalanced Classes

When we experience such issues, we will undoubtedly experience issues settling them with standard calculations. Ordinary calculations are frequently one-sided towards the lion's share class, not thinking about the information dispersion. In the most pessimistic scenario, minority classes are treated as anomalies and overlooked. For a few cases, for example, misrepresentation discovery or disease expectation, we would need to deliberately design our model or falsely balance the dataset, for instance by undersampling or oversampling each class. Notwithstanding, for our situation of learning imbalanced information, the lion's share classes may be of our extraordinary premium. It is alluring to have a classifier that gives high expectation precision over the larger part class, while keeping up sensible exactness for the minority classes. Accordingly, we will abandon it for what it's worth.

E. Text Representation

The classifiers and learning calculations can not straightforwardly process the text archives in their unique shape, as the majority of them expect numerical feature vectors with a settled size instead of the crude text reports with variable length. Along these lines, amid the preprocessing step, the texts are changed over to a more sensible portrayal. One basic methodology for separating features from text is to utilize the pack of words display: a model where for each record, a grievance account for our situation, the nearness (and frequently the recurrence) of words is mulled over, however the request in which they happen is disregarded. In particular, for each term in our dataset, we will compute a measure called Term Frequency, Inverse Document Frequency, abridged to tf-idf. We will utilize "sklearn.feature_extraction.text.TfidfVectorizer" to compute a tf-idf vector for every one of client tweets.

F. Multiclass classifier

To prepare directed classifiers, we initially changed the tweet account into a vector of numbers. We investigated vector portrayals, for example, TF-IDF weighted vectors. In the wake of having this vector portrayals of the text we can prepare administered classifiers to prepare concealed "Customer protestation account" and foresee the "item" on which they fall. After all the above information change, now that we have every one of the features and marks, the time has come to prepare the classifiers. There are various calculations we can use for this sort of issue. Guileless Bayes Classifier: the one most reasonable for word tallies is the multinomial variation.

G. Model Selection

We are presently prepared to try different things with various machine learning models, assess their exactness and discover the wellspring of any potential issues. We will benchmark the accompanying four models, They are Logistic Regression, (Multinomial) Naive Bayes, Linear Support Vector Machine, Random Forest.

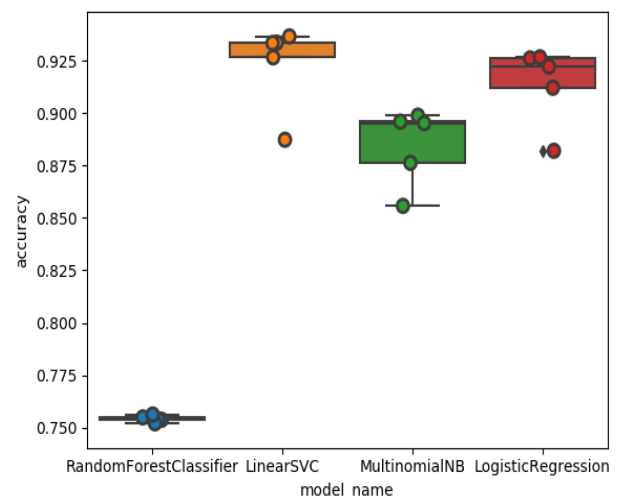


Fig. 1.1 Models Accuracy Graph

IV. DATASET USED

The Dataset used here is CrisisLexT6 which comprises of 2012 sandy hurricane, 2013 Alberta floods, 2013 Boston Bombings, 2013 Oklahoma tornado, 2013 Queensland floods, 2013 Texas Explosion and Random tweets about the incident occurred.

A. Data Integration

Information coordination includes consolidating information living in various sources and furnishing clients with a bound together perspective of them. This procedure winds up critical in an assortment of circumstances, which incorporate both business, (for example, when two comparable organizations need to blend their databases) and logical (joining research results from various bioinformatics archives, for instance) spaces. Information combination shows up with expanding recurrence as the volume (that is, enormous information) and the need to share existing information detonates. It has turned into the focal point of broad hypothetical work, and various open issues stay unsolved.

B. Data Cleansing

Information purging or information cleaning is the way toward distinguishing and revising (or expelling) degenerate or mistaken records from a record set, table, or database and alludes to recognizing inadequate, wrong, off base or unimportant parts of the information and after that supplanting, adjusting, or erasing the grimy or coarse data. Data purifying might be performed intelligently with information wrangling apparatuses, or as clump preparing through scripting. After purging, an informational index ought to be reliable with other comparable informational collections in the framework. The irregularities recognized or evacuated may have been initially caused by client passage mistakes, by defilement in transmission or capacity, or by various information lexicon meanings of comparative elements in various stores. Information cleaning contrasts from information approval in that approval perpetually implies information is rejected from the framework at passage and is performed at the season of section, instead of on bunches of information.

C. Accuracy Achieved

In example acknowledgment, data recovery and double arrangement, exactness (likewise called positive prescient esteem) is the division of pertinent cases among the recovered occurrences, while review (otherwise called affectability) is the portion of significant cases that have been recovered over the aggregate sum of applicable occasions. Both accuracy and review are hence founded on a comprehension and proportion of importance.

TABLE I. ACCURACY OF MODELS

Model No	Accuracy		
	Model Name	Accuracy	Loss
1	Linear SVM	0.82289	0.18821
2	Logistic Regression	0.792297	0.318813
3	Multinomial NB	0.688519	0.422591
4	Random Forest Classifier	0.443826	0.667284

V. RESULT

The dataset is incurred to many classifying models now among those SVM gives the higher accuracy. Hence forth we used the SVM classifier for our model so that we could get the accuracy of 82%

A. Confusion Matrix

Therefore, by finally classifying using the SVM classifier we could able to form a confusion matrix of this kind.

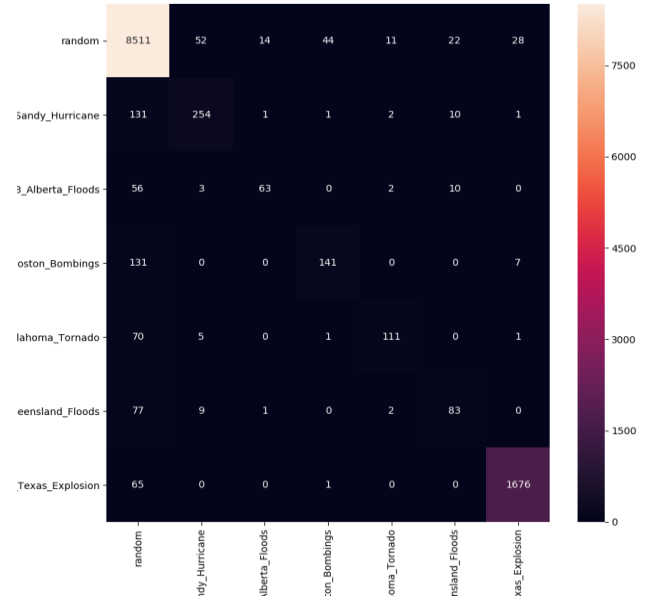


Fig 1.2 Confusion Matrix for labels

B. Output Screenshots

1) After Dataset Preprocessing (represented using pandas)

```

label      tweet      category_id
0      random  I've got enough candles to supply a Mexican fa...      0
1  2012_Sandy_Hurricane  Sandy be soooo mad that she be shattering our ...      1
2      random  @ibexgirl thankfully Hurricane Waugh played it...      0
3      random  @tazs you never got that magnificent case of B...      0
4      random  I'm at Mad River Bar & Grille (New York, N...      0
(35142, 9410)

```

Fig. 1.3 Data Visualization

2) Important Unigrams and Bigrams

```

# '2012_Sandy_Hurricane':
. Most correlated unigrams:
. hurricane
. sandy
. Most correlated bigrams:
. apocalypse hurricane
. hurricane sandy
# '2013_Alberta_Floods':
. Most correlated unigrams:
. abflood
. yycflood
. Most correlated bigrams:
. yyc yycflood
. yycflood http

```

Fig 1.4 Unigrams and Bigrams

3) Output Prediction

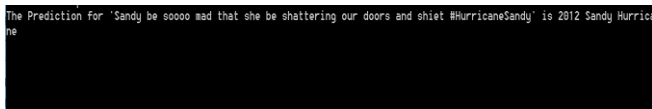


Fig. 1.5 Predicted Result

4) Model Accuracies

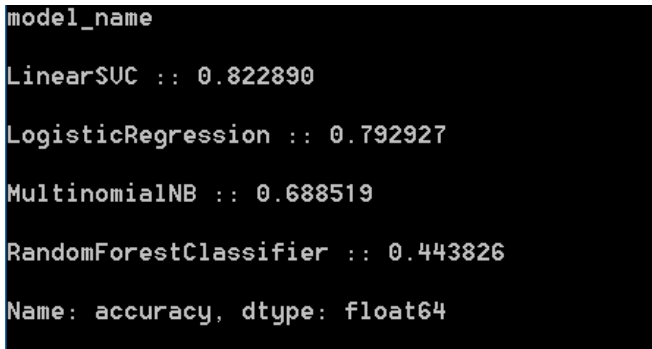


Fig. 1.6 Accuracies of Models Used

VI. CONCLUSION

Thus, we have classified the users interest based on his tweets over the natural disaster event thus we can extend this project to provide the current event news based on the user's interest so that we could enhance the user environment. Hence, this is the thing that we will do is Classifying debacle enthusiasm into 7 predefined classes. Therefore, form categorizing user interest based on his tweets using his tweet w can provide the user interested news over the natural disaster event so that we can provide the user enhanced environment.

VII. REFERENCES

- [1] Rennie, J.D. and Rifkin, R., 2001. Improving multiclass text classification with the support vector machine.
- [2] Ghani, R., 2002, July. Combining labeled and unlabeled data for multiclass text categorization. In *ICML* (Vol. 2, pp. 8-12)
- [3] Chen, J., Huang, H., Tian, S. and Qu, Y., 2009. Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3), pp.5432-5435.
- [4] Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar), pp.1289-1305.
- [5] Forman, G., 2004, July. A pitfall and solution in multi-class feature selection for text classification. In *Proceedings of the twenty-first international conference on Machine learning* (p. 38). ACM.
- [6] Do, C.B. and Ng, A.Y., 2006. Transfer learning for text classification. In *Advances in Neural Information Processing Systems* (pp. 299-306).
- [7] Ko, Y., 2012, August. A study of term weighting schemes using class information for text classification. In *Proceedings of the 35th international ACM SIGIR*

conference on Research and development in information retrieval (pp. 1029-1030). ACM.

- [8] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R. and Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug), pp.1871-1874.
- [9] Uysal, A.K. and Gunal, S., 2012. A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, 36, pp.226-235.
- [10] Wang, T.Y. and Chiang, H.M., 2007. Fuzzy support vector machine for multi-class text categorization. *Information Processing & Management*, 43(4), pp.914-929.