Multi-Label Classification of Parrott's Emotions

Abhijit Mondal, Swapna S. Gokhale
Dept. of Computer Science & Engineering
Univ. of Connecticut, Storrs, CT 06269
{abhijit.mondal,swapna.gokhale}@uconn.edu

Abstract

Mining for latent emotions embedded in tweets can offer clues about users' affective state on a broad range of topics ranging from their mental health to political opinions. This paper presents a multi-class supervised learning approach to group tweets into six emotions (joy, sadness, anger, fear, love, and surprise) defined according to the Parrott's framework. After extensive pre-processing, linguistic and metadata features extracted from a corpus of tweets are used to train popular machine learning classifiers. The performance of these classifiers is evaluated using accuracy, sensitivity, and specificity computed based on a multi-class confusion matrix approach. Our framework can detect common emotions of joy and sadness with excellent accuracy ($> 90\%$), anger and fear with moderate accuracy ($75\% - 85\%$), and love and surprise with lower accuracy ($50\% - 60\%$). Overall, the accuracy of our framework still outperforms that of contemporary approaches for all the six emotions. Further analysis of an example multi-class confusion matrix indicates that lower accuracy values for love and surprise may arise because love is often confused with joy, whereas surprise is mixed up with the positive emotion of joy and the negative emotion of fear. Moreover, this confusion could be attributed to an under-representation of these emotions in the data. This highlights the need for building high-quality, balanced benchmark data sets for training multi-label emotion classifiers.

1 Introduction

Social media platforms such as Twitter, Facebook and Instagram offer a forum for people to share and communicate with large audiences as they go about their daily lives. Twitter is one of the most popular social media platforms, with nearly 330 million monthly active users on an average as of 2019 [3]. Twitter's large, active user base generates volumes of textual content in the form of tweets. This content shared by the users is interactive, spontaneous, conversational, and unfiltered. Tweets thus contain a treasure trove of information that can offer clues about users' opinions, thoughts, and feelings on a variety of topics from politics to restaurants to even their mental health.

The plethora of information embedded in these tweets has attracted significant attention in their mining and analysis. A large body of work has focused on detecting and classifying the sentiment and/or polarity of the tweets [21]. In binary sentiment analysis, tweets are grouped according to positive and negative polarities, whereas in multi-class analysis they are grouped into more than two classes according to the strength of the embedded sentiment. Tweets, however, also contain affective information (moods, emotions, and feelings) of the users, and they can also be mined for these emotions. Emotion mining can thus be viewed as a deeper, more advanced form of sentiment analysis [12]. This detailed, granular information that can be extracted from tweets can support a range of applications such as targeted advertising, recommending books, music and videos, predicting the movements of stock markets, launching television programs, detecting and monitoring mental health problems, and gathering public opinion on politically and socially sensitive issues.

Emotion classification can be binary, where opposing emotions such as joy and sadness or love and hate are formulated into targeted two-way detection problems. Binary emotion classification problems can also be formulated by combining all the positive emotions such as love, joy, and trust into one class, and all the negative emotions including hate, sadness and disgust into another class. Plutchik's wheel provides a natural anchor for formulating such two-way problems, as opposing emotions are placed on the two opposite ends of each axis on a wheel [11]. Multi-label classification of emotions, on the other hand, involves grouping tweets into many classes; these classes are usually chosen in a manner that is convenient based on the data, or are in some cases inspired by a psychological framework such as the Plutchik's wheel [15], the Parrott's framework [13] or

DOI reference number: 10.18293/SEKE2021-158

the Ekman's atlas of emotions [6]. Overall, in the literature, multi-label classification shows lower accuracy for all the classes or is seen to trade away the accuracy of one class for the other [14, 9, 19]. This could occur because all the emotions in a multi-class problem may not be expressed to a similar degree, that is, the data could be unbalanced. Another reason could be that these uncommon emotions are often confused or mistaken for the commonly occurring ones. To the best of our knowledge, other than the fact that multi-label emotion detection is a challenging problem, very little is known in the way of reasons behind the challenge. This objective of this paper is to present a framework that can classify a corpus of tweets into multiple emotions with good accuracy over contemporary approaches. A secondary objective is to gain deeper insights into the challenges involved in building high accuracy multi-label classifiers through a more in-depth analysis. The approach is built around a recently annotated data set [16], which tags each tweet with one of six emotions. We map these six labels to the six basic emotions defined by the Parrott's model [13]. We extensively pre-process these tweets, extract linguistic and metadata features, and train five popular machine learning models using these features. We evaluate the performance of these models using accuracy, sensitivity, and specificity, computed based on the multi-class confusion matrix approach.

Our results indicate that the more basic and common emotions of joy and sadness can be identified with excellent accuracy ($> 90\%$), anger and fear with moderate accuracy ($75\% - 85\%$), and love and surprise with low accuracy ($50\% - 60\%$). With these accuracy values, our classifiers still perform better than the current approaches for all the emotions. The classifiers show higher specificity compared to sensitivity, which means that they are better at ruling out a specific emotion rather than identifying it affirmatively. An analysis of an example multi-class confusion matrix indicates that love is often confused with joy, whereas surprise is mixed up with the positive emotion of joy and the negative emotion of fear, which could explain the low detection accuracy for these emotions. This confusion could occur because love and surprise are complex emotions which embody both positive and negative feelings. Moreover, because of their complexity, these emotions could be underrepresented in our corpus compared to the other classes; especially joy and sadness. Therefore, one of the ways in which the accuracy of multi-label emotion classification may be improved is by building high-quality training data sets, with a balanced representation of all the involved emotions.

The rest of the paper is organized as follows: Section 2 presents the emotion classification model. Section 3 describes the steps in the classification framework. Section 4 discusses the results. Section 5 compares and contrasts related research. Section 6 concludes the paper and offers directions for future research.

2 Emotion Classification Model

We used the data set made available by Saravia *et al* [16]. This set of tweets was collected using a set of hashtags, which served as noisy labels for subsequent distance-based annotations. Each tweet is labeled into one of six emotions: joy, sadness, anger, fear, love, and surprise. Annotated tweets were split into train and test data sets, with the total number of tweets in these partitions being 16000 and 2000 respectively. The number and percentage of tweets in the train/test partitions for the six emotions are summarized in Table 1. The table shows the imbalance between the emotions; joy and sadness are the most common; fear, anger and love form the next tier; whereas surprise is the most rare. However, the train/test split was conducted using stratified sampling because the ratio of test to train is maintained between 11% and 13% for all the emotions.

Emotion	Train		Test		Test/Train
	#	%	#	%	
Joy	5362	33.5%	695	34.75%	13.00%
Sadness	4666	29.17%	591	29.55%	12.67%
Fear	1937	12.10%	224	11.20%	11.57%
Anger	2159	13.50%	275	13.75%	12.73%
Love	1304	8.15%	159	8.00%	12.19%
Surprise	572	3.58%	66	3.3%	11.54%

Table 1: Summary of Tweets Per Emotion

We referred to the three most popular models of emotions to formulate the multi-label classification problem. These are the Gerrod Parrott's model containing six basic human emotions [13], the Plutchik's wheel of emotions [15], and the Ekman's atlas of emotions [6]. The emotion labels in our data coincide exactly with Parrott's model, providing us a natural anchor for our six-way classification problem.

3 Classification Framework

This section describes the classification framework.

3.1 Data Pre-processing

Our data consisted of the text of the tweets and its emotion label. It was relatively clean, and there were no emoticons, punctuations, links, hashtags and other markers. So, the pre-processing steps were relatively straightforward. First we tokenized all the words using white spaces.

In the second step, we removed the stop words using the stop words list in NLTK library [18]. In the third step, we removed proper nouns (names of persons, cities, etc). and other non-English words by checking the presence of every word against the NLTK list.

3.2 Feature Extraction

We considered two linguistic features: Term Frequency-Inverse Document Frequency (TF-IDF) vectorization and n-grams using the bag-of-words approach. These features were extracted using the TF-IDF vectorizer class of Scikit-Learn library [2]. TF-IDF refers to a scoring measure used in information retrieval or summarization. It measures the relevance of a word in a document by assigning an additional weight to frequent words. We computed the TF-IDF scores for the topmost 1000 unigrams.

We extracted six meta-data features from the text of each tweet prior to pre-processing. These include the number of characters, number of words, number of stop words, and number of unique words, TextBlob and Vader sentiment scores. TextBlob calculates the sentiment polarity for each tweet, which ranges from -1 to $+1$, where $-1, 0$ and $+1$ indicate negative, neutral and positive respectively. Vader computes a compound score as a normalized and weighted composite score obtained by analyzing each word in a tweet for its direction of sentiment – a negative (positive) valency for negative (positive) sentiment. It therefore ranges from -1 to $+1$ depending on the net sentiment of the tweet. We used both TextBlob and Vader scores because Vader may be more sensitive to sentiments than TextBlob, even though TextBlob may be better correlated with reviewer scores [1].

3.3 ML Models

We employed the following common machine learning models for classification. Implementations of these models in the Scikit-Learn and Keras libraries were used.

- **Random Forests (RF):** Random Forests is an ensemble learning classification technique based on decision trees. The number of decision trees is set to 30 and the number of features used by each tree is equal to the squared-root of the number of total features. Finally, each tree was allowed to grow fully up to its leaves.

- **Support Vector Machines (SVM):** Support Vector Machines is a classification method that estimates the boundary (called hyper-plane) with the maximum margin. We used SVMs with linear kernel with other default parameters.

- **Multi-Layer Perceptron (MLP):** Multi-layer Perceptron is a deep neural network that consists of input, hidden, and output layers. Our MLP model consisted of 3 hidden layers with 10, 5, and 2 neurons respectively, along with the rectifier linear unit (ReLu) activation function.

- **Gradient Boosting (GB):** Gradient Boosting is another ensemble learning classifier which builds classifier trees such that each tree takes a small step towards the minimization of classification error from the previous tree. The algorithm continues until maximum number of trees are built or there is no significant improvement in minimizing the error. Finally, predictions for the test data are obtained by combining predictions of the trees built in each stage using a weighted sum. We used 100 estimators, with a maximum depth of 1.

- **Neural Network (NN):** We build a neural network with three layers having 30, 10 and 6 neurons respectively. We arrived at this architecture through experimentation, considering that our data was of medium complexity with about 1000 features.

3.4 Performance Metrics

For multi-label classification, the first step in defining the performance metrics is the computation of the multi-class confusion matrix, which represents how many of the tweets originally in that class are classified accurately as belonging to that class. Also, for a given class it represents the number of tweets that are mis-labeled by a classifier as belonging to each of the other five classes. Finally, we divided each of these six counts by their sum to obtain a normalized accuracy measure. For example, let 500 tweets be originally labelled as "surprise". Now, suppose if 400 of these tweets were labelled correctly by the classifier as "surprise" but the other 100 tweets were mis-labelled. Further, suppose that these 100 tweets were mis-labelled equally among the other five classes meaning each of the other five classes included 20 of these tweets. Next, we divide these six counts by 500 to compute the six elements in the normalized multi-class confusion matrix as 0.8, 0.04, 0.04, 0.04, 0.04, and 0.04. We repeat this process to calculate all 36 entries in the confusion matrix. The accuracy for each class is defined as the percentage of tweets labeled correctly from that class, and refers to the diagonal elements in the confusion matrix.

Alongside multi-label classification accuracy, we also calculated two other performance metrics, namely, sensitivity and specificity. Sensitivity and specificity together offer insights into the bias of a classifier towards a particular class. However, these two performance metrics are mainly used in the context of binary classification problems, as they need us to define positive and negative classes in order to be able to compute true and false positives, and true and false

negatives. Therefore, we transformed this multi-label classification problem as six "one vs rest" classification problems. For example, to calculate the sensitivity and specificity for surprise, we considered the "surprise vs rest" classification problem. We designated the positive class as "surprise" and all the other classes together formed the negative class. True positives (TP) are the tweets which are correctly classified as "surprise", true negatives (TN) include tweets originally not from the "surprise" class and are also not labeled as "surprise" by a classifier. Similarly we can define false positives (FP) as those tweets that were incorrectly labeled as "surprise", and false negatives (FN) as those tweets that were originally labeled as "surprise" but the classifier labeled them incorrectly with one of the other five classes.

Equation (1) shows the expressions for sensitivity and specificity. Sensitivity of "surprise" class is the percentage that a tweet labeled as "surprise" is correctly classified as such. We note that sensitivity is identical to multi-label accuracy. If a highly sensitive classifier classifies a tweet into an emotion class, then it can be fairly certain that it actually does. Specificity of "surprise" class is the percentage that a tweet which is not labeled as "surprise" is classified as such. If a highly specific classifier says that the tweet does not exhibit an emotion, then we can be fairly certain that it indeed does not. Generally, there is a trade-off between sensitivity and specificity. A classifier with a high sensitivity usually has low specificity, and vice versa.

After computing the sensitivity/accuracy and specificity for each emotion, we compute the aggregate unweighted and weighted values of these metrics across all classes. The weight for each class is given by the percentage of tweets in that class in the training data set.

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{TN+FP} \end{aligned} \tag{1}$$

4 Results & Discussion

The data was already split into train and test sets for model training and performance evaluation. Table 2 shows the per-class accuracy for the five models. With our unweighted and weighted overall accuracy values of 74% and 83% for the NN model, our approach outperforms many contemporary approaches with accuracy values in the ranges of $50-60\%$ [4, 19, 9].

Tables 2 and 3 summarize the sensitivity/accuracy and specificity values for the six classes by the five classifiers. Across all the emotions, sensitivity values are lower than their corresponding specificity values. All specificity values are around or over 95%, meaning that the classifiers are excellent at negative differentiation, that is, they can identify with near certainty, the absence of a specific emotion. Lower sensitivity values indicate that the classifiers are less capable of zeroing in on a specific emotion. The classifiers can identify joy and sadness with excellent accuracy, fear and anger with moderate to low moderate accuracy, but struggle with love and surprise; more so with surprise than with love. Albeit low, our accuracy in detecting surprise still exceeds the accuracy of the contemporary works that have simply been unable to detect this emotion [14]. Neural networks offer the best specificity across all the emotions. Sensitivity produces mixed results among models; for each emotion the best model for sensitivity is different and is identified in the parentheses: joy (SVM), sadness and fear (NN), anger and love (MLP), and surprise (RF). Generally, the difference in sensitivity/accuracy between the models is small for all emotions except for surprise, where MLP diverges significantly.

To understand why the classifiers may struggle with the emotions of love and surprise, we take a closer look at an example multi-class confusion matrix from the SVM model (matrices from other models show similar trends) as shown in he Figure 1. This matrix shows how the tweets from each class are mis-classified into the other five classes. From the figure, it can be seen that love is most likely to be confused with joy while surprise is most likely to be confused with either joy or fear. Therefore, the expression of love is almost always positive, whereas, surprise can be expressed in both positive and negative senses; and it embodies both these emotions. This confusion, which leads to lower accuracy for love and surprise could be due to class imbalance; Table 1 shows that only 8% tweets are labeled as love, and an even lower 3% tweets are labeled as surprise.

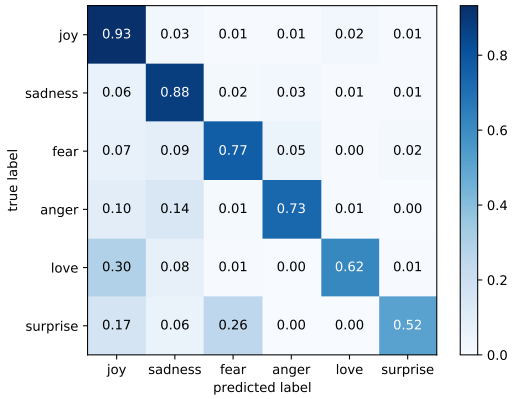

Figure 1: Multiclass confusion matrix (SVM)

Model	Joy	Sadness	fear	Anger	Love	Surprise	Unweighted	Weighted
NN	0.91	0.91	0.84	0.75	0.64	0.53	0.74	0.83
RF	0.87	0.82	0.71	0.68	0.50	0.64	0.68	0.76
SVM	0.93	0.88	0.77	0.73	0.62	0.52	0.74	0.83
MLP	0.84	0.88	0.79	0.78	0.76	0.17	0.77	0.67
GB	0.84	0.79	0.76	0.77	0.61	0.46	0.71	0.78

Table 2: Sensitivity/Accuracy of ML Classifiers

Model	Joy	Sadness	fear	Anger	Love	Surprise	Unweighted	Weighted
NN	0.95	0.96	0.94	0.98	0.98	0.99	0.97	0.96
RF	0.84	0.89	0.99	0.99	0.99	0.99	0.95	0.91
SVM	0.90	0.93	0.98	0.98	0.99	0.99	0.96	0.94
MLP	0.95	0.94	0.96	0.96	0.96	0.99	0.96	0.95
GB	0.86	0.89	0.98	0.98	0.99	0.99	0.95	0.91

Table 3: Specificity of ML Classifiers

5 Related Research

Prevalent research efforts have mined emotions surrounding specific events such as the presidential election [20] or the Brazilian soccer league [5], or natural disasters such as the California Camp Fire [10] and the MERS outbreak [4]. However, extracting them from a general corpus remains relatively unaddressed.

Many research works formulate multi-label classification problems over a set of emotions; the chosen set may be completely ad hoc, inspired by a psychological framework such as the Ekman's atlas of emotions [6] or the Plutchik's wheel [15], or a combination of psychology and heuristics. For example, Wang *et. al.* [19] annotated a data set of 2.5 million tweets based on hashtags related to emotion words, and classified them into seven emotions, six basic plus "thankfulness". Their classification accuracy is around 60%, and this performance is further improved by about 5% [9]. Jaishree et. al. [14] label tweets by combining the scores from NRC word-level lexicon tool and emotion-based hashtags. Their problem considered 8 basic emotions on the Plutchik's wheel, however, their multi-label classification problem was completely unable to detect surprise, and registered low scores for fear. A smaller set of 4 emotions is also used by some [7, 17]. Although Mohammed *et. al.* formulate their problem based on the Plutchik's wheel, they ultimately boil it down to binary classification by using the one vs. other method [8]. Generally, multi-label emotion classification suffers from either low accuracy for all classes or sacrifice the accuracy of some for the others. The accuracy values of our approach are higher for all the emotions compared to these contemporary approaches. Moreover, a detailed analysis sheds further light into those emotions that are difficult to detect, and how they could be confused with the others.

6 Conclusions and Future Research

Simultaneous differentiation between multiple emotions from content shared on social media platforms remains a challenging problem. This paper proposes a classification framework based on supervised machine learning that can identify six emotions of joy, sadness, anger, fear, love, and surprise defined in the Parrott's framework from a corpus of tweets. Relying on extensive pre-processing of tweets, followed by the extraction of linguistic and metadata features to train popular machine learning models, our classification framework can identify joy and sadness with excellent accuracy, anger and fear with moderate accuracy, and love and surprise with low accuracy. Moreover, the aggregate accuracy of our approach is better than contemporary approaches. Through a detailed analysis, we develop insights into why love and surprise could be difficult to detect, and offer that one plausible explanation for this difficulty could stem from an under-representation of these two emotions in the data.

Our future research involves building a high-quality balanced data set that can be used to train classifiers for multi-label emotion classification. Experimenting with identifying emotions surrounding high profile events related to Covid-19 such as vaccinations, or the passage of the American Rescue Plan Act of 2021 is also a topic of the future.

References

[1] K. Arunachalam. *"Evaluation of Python Packages for Sentiment Analysis"*, 2019 (last accessed on October 19, 2019). https://www.linkedin.com/pulse/evaluation-python-packages-sentiment-analysis-karthikeyan-arunachalam/.

[2] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and Gaël Varoquaux. "API Design for Machine Learning Software: Experiences from the Scikit-learn project". In *Proc. of ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[3] J. Clement. *"Number of Monthly Active Twitter Users Worldwide from 1st Quarter 2010 to 1st Quarter 2019"*, 2019 (last accessed on March 5, 2020). https://www.statista.com/statistics/282087/.

[4] H. J. Do, C. Lim, Y. J. Kim, and H. Choi. "Analyzing Emotions in Twitter during a Crisis: A Case Study of the 2015 Middle East Respiratory Syndrome Outbreak in Korea". In *Proc. of Intl. Conf. on Big Data and Smart Computing*, pages 415–418, 2016.

[5] A. Esmin, R. De Oliveira Jr, and S. Matwin. "Hierarchical Classification Approach to Emotion Recognition in Twitter". In *Proc. of Intl. Conf. on Machine Learning and Applications*, volume 2, pages 381–385, 2012.

[6] S. Handel. "Classification of Emotions". https://www.theemotionmachine.com/article-limit/, 2011.

[7] S. S. Ibraheim, S. S. Ismail, K. A. Bahansy, and M. M. Aref. "Multi-Emotion Classification Evaluation via Twitter". In *Proc. of Intl. Conf. on Intelligent Computing and Information Systems*, pages 60–67, Cairo, Egypt, 2019.

[8] M. Jabreel and A. Moreno. "A Deep Learning-based Approach for Multi-label Emotion Classification in Tweets". *Applied Sciences*, 9(6):1123, 2019.

[9] O. Janssens, M. Slembrouck, S. Verstockt, S. V. Hoecke, and R. V. de Walle. "Real-time Emotion Classification of Tweets". In *Proc. of Intl. Conf. on Advances in Social Network Analysis and Mining*, pages 1430–1431, August 2013.

[10] N. H. Khun, T. T. Zin, M. Yokota, and H. Y. Thant. "Emotion Analysis of Twitter Users on Natural Disasters". In *Proc. of Global Conf. on Consumer Electronics*, Osaka, Japan, October 2019.

[11] Abhijit Mondal and Swapna S. Gokhale. Mining emotions on plutchik's wheel. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–6, 2020.

[12] R. M. Ohashi. *"From Sentiment Analysis to Emotion Recognition: A NLP Story"*, July 2019 (last accessed on March 3, 2020). https://medium.com/neuronio/bcc9d6ff61ae.

[13] G. W. Parrott. *Emotions in Social Psychology: Essential Readings*. Psychology Press, 2001.

[14] J. Ranganathan, N. Hedge, A. S. Irudayaraj, and A. A. Tzacheva. "Automatic Detection of Emotions in Twitter Data-A Scalable Decision Tree Classification Method". In *Proc. of the RevOpID 2018 Workshop on Opinion Mining, Summarization and Diversification*, 2018.

[15] Plutchik Robert. Emotion: Theory, research, and experience. vol. 1: Theories of emotion, 1980.

[16] E. Saravia, H. Liu, Y. Huang, and Y. Chen. "CARER: Contextualized Affect Representations for Emotion Recognition". In *Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, November 2018.

[17] F. M. Shah, A. S. Reyadh, A. I. Shaafi, and F. T. Sithil. "Emotion Detection from Tweets using AIT-2018 Dataset". In *Proc. of Intl. Conf, on Advances in Electrical Engineering*, Dhaka, Bangladesh, September 2019.

[18] Bird Steven and L Edward. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72, 2006.

[19] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. "Harnessing Twitter "Big Data" for Automatic Emotion Identification". In *Proc. of Intl. Conf. on Privacy, Security, Risk and Trust and Intl. Conf. on Social Computing*, pages 587–592, 2012.

[20] U. Yaqub, S. Chun, V. Atluri, and J. Vaidya. "Sentiment-based Analysis of Tweets during the US Presidential Election". pages 1–10, 06 2017.

[21] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin. "A Survey of Sentiment Analysis in Social Media". *Knowledge and Information Systems*, (60):617–663, July 2018.