ISSN 2007-9737

Multi-label Emotion Classification using Content-Based Features in Twitter

Iqra Ameer[1], Noman Ashraf[1], Grigori Sidorov[1], Helena Gómez Adorno[2]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

[2] Universidad Nacional Autónoma de México,
Instituto de Investigación en Matemáticas Aplicadas y en Sistemas,
Mexico

iqraameer133@gmail.com, nomanashraf@sagitario.cic.ipn.mx,
sidorov@cic.ipn.mx, helena.gomez@iimas.unam.mx

Abstract. Multi-label Emotion Classification is a supervised classification problem that aims to classify multiple emotion labels from a given text. Recently, Multi-label Emotion Classification has appealed to the research community due to possible applications in E-learning, marketing, education, and health care, etc. We applied content-based methods (words and character n-grams) on tweets to show how our purposed content-based method can be used for the development and evaluation of the Multi-label Emotion Classification task. The results achieved after our extensive experimentation demonstrate that content-based word unigram surpassed other content-based features (Multi-label Accuracy = 0.452, $Micro_{F_1} = 0.573$, $Macro_{F_1} = 0.559$, Exact Match = 0.141, Hamming Loss = 0.179).

Keywords. Multi-label emotion classification, content-based methods, twitter.

1 Introduction

Multi-label Emotion Classification (MEC) is a supervised classification task to determine the presence of multiple emotions in a given piece of text. Single-label Emotion Classification captures only one emotion in text whereas MEC captures all present emotions which best represent the mental state of the writer. MEC has captured the intention of the research community due to its potential applications in various domains including, E-learning, marketing, education, health care [15], etc.

Due to the current boom in data creation and the usage of smartphones by people which results in them expressing their thoughts over public social media communication forums, such as Facebook, Twitter, YouTube, etc. There is no doubt that humans are emotional creatures, emotions are an important part of human life, and affect our choices as well as our psychological and physical health. In this case, the implementation of MEC systems is necessary. There are two types of automatic Emotion Classification: (1) Single-label Automatic Emotion Classification — associate a single label to the given instance from the finite set of pre-defined labels, which best describes the emotion in the given instance, and (2) Multi-label Automatic Emotion Classification — associate multiple labels to the given instance from the finite set of pre-defined labels, which best describe the mental state of the author.

MEC on Twitter is a challenging task as only 280[1] characters are allowed and users have to use informal language structure such as a short form of texts, emojis' etc. Another reason is a tweet may contain more than one emotion label in a given content.

This study describes our development process of the automatic MEC model for tweets. We used content-based

[1] https://blog.twitter.com/en_us/topics/product/2017/Giving\protect\discretionary{\char\hyphenchar\font}{}{}you-more-characters-to-express-yourself.html

Table 1. Example of Multi-label Tweet Dataset

No.	Tweet	Emotions
01	@NHLexpertpicks @usahockey USA was embarrassing to watch. When was the last time you guys won a game..? #horrible #joke	anger, disgust
02	My sister is graduated with 3.9 CGPA	happy, love, surprise

methods particularly word n-grams and character n-grams as features along with their combinations. We used MEKA[2] implementation of four multi-label and 11 single-label classifiers. Methods were evaluated by using the tweets dataset of SemEval-2018[3] competition. The remaining article is structured as follows: Section 2 represents the related work. Section 3 briefs the task description and corpus provided by the SemEval-2018. Section 4 describes the evaluation methodology, Section 5 concludes the article and proposes future work.

2 Related Work

Lately, scientists have shown significant attention to MEC in the textual content. In this section, we explore the prior work of this domain. The A-A is an unsupervised emotion classification model [16] depends on rules and a manually labeled corpus. The model comprises emoticons' words with affects, acronyms, and familiar abbreviations. EC-VSM [8] is also an unsupervised cosine similarity-based model, unigram features weighted by TF-IDF and improved by lexica like WordNet Affect Lexicon [20]. The unsupervised model introduced in [11] utilizes reduction tools and lexicon, for example, Non-negative Matrix Factorization (NNMF) and Latent Semantic Analysis (LSA).

As opposed to these techniques, supervised-learning has been merged with a psychological methodology as expressed in [9]. Specifically, a Hidden Markov Model (HMM) was utilized to reproduce how mental state sequences influence or cause emotions. [6] removed the stop words and generated the features by TF-IDF weightage and SenticNet lexicon.

[14] utilized unigram and bi-grams of words emotion classification task on newspaper headlines' corpus. [13] used the unigram and bi-grams of words along with elongated words, punctuations marks, emotion-related lexicons and negation features to classify the emotional state and the stimulus on a Twitter corpus of 2012 US presidential elections. [14] utilized features associated with emotions, word n-grams, and elongated words for the emotion detection task.

Single-label classifier like Support Vector Machine (SVM) that characterized the content using unigrams used with a multi-label classifier such as Label Powerset (LP) to detect emotions from suicide notes [10] and 15 emotions detected such as anger, joy, guilt, and love, etc. To detect emotions from Brazilian Portuguese short texts, a lot of multi-label classifiers used for example RAkEL and HOMER [1].

The Semantic Evaluation series (SemEval-2018)[4] played an important role in the emotion Classification task. In task 4 SemEval-2007 competition, the organizers provided the news headlines and asked to classify the polarity and emotions [19]. The rule-based system UPAR7 that uses dependency graphs performed best out of all three participants [7]. In task 5 SemEval-2018 contest, the Maximum number of researchers used lexica, word n-grams or word embeddings along with Deep Learning (DL) based models such as Convolution Neural Network (CNN), Recurrent Neural Network (RNN) or Long-Short Term Memory Network (LSTM) architectures to classify multiple emotions from the text. The best performing team (NTUA-SLP) implemented Bidirectional LSTM (Bi-LSTM) with a multilayer self-attention mechanism [12], as Bi-LSTM used to perform well in classification task [2].

3 Task Description and Dataset

MEC on Twitter at SemEval-2018 had corpora for the Arabic, English, and Spanish languages. However, for this study, we have chosen only the English language.

3.1 Task Description

The MEC task: A tweet is given, determine the tweet as "neutral or no emotion" or one, or more from a set of 12 following emotion labels that best show the emotional condition of the author:

— Anger (also frustration and wrath),

[2] http://waikato.github.io/meka/

[3] https://competitions.codalab.org/competitions/17751

[4] https://semeval.github.io/

Table 2. Percentage of Tweets that were annotated with a given emotion

Language	anger	anti.	disg.	fear	joy	love	optm.	pessi.	sadn.	surp.	trust	neutral
Eng	36.1	13.9	36.6	16.8	39.3	12.3	31.3	11.6	29.4	5.2	5.0	2.7

— Anticipation (also concern and attention),
— Disgust (also disregard and dislike),
— Fear (also nervousness, worry, and panic),
— Joy (also peacefulness and delight),
— Love (also affection),
— Optimism (also hope and self-assurance),
— Pessimism (also distrust, no assurance),
— Sadness (also sorrow, unhappiness),
— Surprise (also shock, amazement),
— Trust (also liking, approval, and acceptance).
— Neutral or no emotion.

3.2 Dataset

The MEC in tweets (ML-EC-2018) Dataset is collected to verify the presence/absence of 11 emotions. The dataset includes a total of 10,983 tweets.

The tweets were divided into train, dev, and test groups with 6,838 tweets in the training set, 886 tweets in Dev, 3,259 tweets in the test set. Table 2 shows the percentage of tweets for each emotion label.

Please note, the sum is more than 100% because a tweet possibly annotated with more than one emotion type.

4 Evaluation Methodology

The MEC task is treated as a supervised classification problem. The goal is to classify tweets as of one, none or more of 12 emotions that best portrays the emotional state of the author. We used 10-fold cross-validation to better estimate the performance of the content-based method.

This section represents our methodology considering pre-processing, evaluation measures used in this study, the features set, single-label, and multi-label machine learning (ML) algorithms used for the MEC problem.

4.1 Pre-processing

Before feature extraction process, we applied following pre-processing steps [3, 18] on dataset of SemEval-2018:

— Lower-cased the tweets,
— Punctuation marks are removed,
— Stop words are removed,
— Normalized the elongated words.

4.2 Evaluation Measures

To evaluate this study, we used official evaluation measures of SemEval-2018 task: (i) Accuracy (multi-label), (ii) Micro-averaged F_1 score, (iii) Macro-averaged F_1 score[5], along with official measure, we also reported results with: (i) Exact Match, (ii) Hamming Loss:

— **Exact Match:** Calculates the percentage of samples whose predicted labels are the same as their gold labels.
— **Hamming Loss:** Calculates the average of how many times a label of an instance is classified incorrectly.

4.3 Features

The pre-processed text was used to generate the features for single-label and multi-label ML algorithms. In this study, we applied Content Based Method, which aims to classify the author's emotions by analysis content of the message. We applied 10 content based features – 3 word-based features and 7 character-based features.

We also performed experiments with combinations of word n-grams and combination for character n-grams as they supposed to improve the results in classification problems [4, 17].*N* varies from 1-3 for word-based features and 3-9 for character-based features. We weighted the features by well-known TF-IDF scores by using Scikit-learn[6] implementation.

[5] https://competitions.codalab.org/competitions/17751

[6] https://scikit-learn.org/

Table 3. Best results obtained using content based methods

Features	MLC	SLC	Acc.	EM	HL	Mi-F_1	Ma-F_1
Word 1-gram	**BR**	**RF**	**0.452**	**0.141**	**0.179**	**0.573**	**0.559**
Word 2-gram	BR	DT	0.366	0.053	0.200	0.515	0.496
Word 3-gram	LC	SMO	0.308	0.136	0.237	0.373	0.363
Character N-grams							
Char 3-gram	**CC**	**Bagging**	**0.354**	**0.347**	**0.117**	**0.340**	**0.357**
Char 4-gram	CC	SMO	0.334	0.330	0.124	0.313	0.336
Char 5-gram	BR	Bagging	0.329	0.294	0.132	0.320	0.342
Char 6-gram	CC/LC	ASC/FC/DT/J48	0.331	0.331	0.123	0.310	0.331
Char 7-gram	LC	DT	0.335	0.335	0.125	0.313	0.335
Char 8-gram	LC	DT	0.335	0.335	0.125	0.313	0.335
Char 9-gram	LC	DT	0.335	0.335	0.125	0.313	0.335
Combination of Word N-grams							
Word 1-3-gram	BR	RF	0.451	0.137	0.179	0.572	0.558
Combination of Character N-grams							
Char 3-9-gram	BR	RF	0.287	0.012	0.269	0.406	0.396

4.4 Single-label and Multi-label Machine Learning Algorithms

We have an MEC task, and the aim is to classify multiple emotions among 12 emotions. In this paper, we tried MEKA implementation of various machine learning multi-label along with single-label classifiers however, we are reporting results with only best-performing algorithms. Multi-label classifiers including Binary Relevance (BR), BPNN, Classifier Chain (CC), and Label Combination (LC) and single-label classifiers involving BayesNet, SGD, SMO, Voted Perceptron, AdaBoostM1, AttributeSelectedClassifier(ASC), Bagging, FilteredClassifier(FC), DecisionTable(DT), J48, and RandomForest(RF).

5 Results and Analysis

In Table 3, only the best scores are represented. From Table 3, we can notice that the best results are achieved on word-unigram feature by using multi-label Binary Relevance and single-label Random Forest classifiers (Multi-label Accuracy = 0.452, $Micro_{F_1} = 0.573$, $Macro_{F_1} = 0.559$, Exact Match = 0.141, Hamming Loss = 0.179). The combination of word N-grams (n = 1-3) also performing the same as word unigrams with a slight difference of 0.001.

Character n-grams are not performing well to classify emotions. Therefore we can say that, for all considered evaluation measures, word 1-gram (content-based feature), together with Binary Relevance and Random Forest classifiers is helpful in the MEC problem.

For multi-label classifiers, Binary Relevance is performing better as compared to the other four multi-label classifiers. The behavior of a single-label classifier is fluctuating with the combination of different multi-label classifiers. This indicates that the performance of the Binary Relevance is dependent on a single-label machine learning classifier or vice versa.

6 Conclusion and Future Work

In this article, we tackled MEC as a supervised classification problem. We considered 12 emotion labels as in SemEval-2018 Task-1 (see section 3.1). We implemented content-based methods (words and character n-grams) and performed extensive experimentation. Results show that word unigrams are better as a feature to classify multiple emotions from a given tweet (Multi-label Accuracy = 0.452, $Micro_{F_1} = 0.573$, $Macro_{F_1} = 0.559$, Exact Match = 0.141, Hamming Loss = 0.179).

To handle the MEC problem, the combination of multi-label Binary Relevance with a single-label Random

Forest classifier is best. The possible future work is as follows: we plan to explore neutral deep learning for the emotion classification task, for example, Bi-LSTM, attention mechanism, self-attention.

We will also see how stylometry-based methods behave on an emotion classification problem. We can also increase the dataset by using data augmentation techniques to see the behaviors of different classification methods on the MEC problem [5].

Acknowledgements

This work is done with partial support of CONACYT, SNI, CONACYT project A1-S-47854, SIP projects 20200797 and 20200859.

References

1. **Almeida, A., Cerri, R., Paraiso, E. C., Mantovani, R. G., & Junior, S. B. (2018).** Applying multi-label techniques in emotion identification of short texts. *Neurocomputing*, Vol. 320, pp. 35–46.
2. **Amajd, M., Kaimuldenov, Z., & Voronkov, I. (2017).** Text classification with deep neural networks. *International Conference on Actual Problems of System and Software Engineering (APSSE)*, pp. 364–370.
3. **Ameer, I., Siddiqui, M. H. F., Sidorov, G., & Gelbukh, A. (2019).** CIC at SemEval-2019 task 5: Simple yet very efficient approach to hate speech detection, aggressive behavior detection, and target classification in twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 382–386.
4. **Ameer, I., Sidorov, G., & Nawab, R. M. A. (2019).** Author profiling for age and gender using combinations of features of various types. *Journal of Intelligent & Fuzzy Systems*, Vol. 36, No. 5, pp. 4833–4843.
5. **Amjad, M., Sidorov, G., & Zhila, A. (2020).** Data augmentation using machine translation for fake news detection in the urdu language. *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2537–2542.
6. **Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018).** SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Thirty-Second AAAI Conference on Artificial Intelligence*.
7. **Chaumartin, F.-R. (2007).** UPAR7: A knowledge-based system for headline sentiment tagging.
8. **Danisman, T. & Alpkocak, A. (2008).** Feeler: Emotion classification of text using vector space model. *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, pp. 53.
9. **Ho, D. T. & Cao, T. H. (2012).** A high-order hidden Markov model for emotion detection from textual data. *Pacific Rim Knowledge Acquisition Workshop*, Springer, pp. 94–105.
10. **Luyckx, K., Vaassen, F., Peersman, C., & Daelemans, W. (2012).** Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. *Biomedical informatics insights*, Vol. 5, pp. BII–S8966.
11. **Mac Kim, S., Valitutti, A., & Calvo, R. A. (2010).** Evaluation of unsupervised emotion models to textual affect recognition. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 62–70.
12. **Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018).** Semeval-2018 task 1: Affect in tweets. *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17.
13. **Mohammad, S., Zhu, X., & Martin, J. (2014).** Semantic role labeling of emotions in tweets. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 32–41.
14. **Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015).** Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, Vol. 51, No. 4, pp. 480–499.
15. **Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B., & Gelbukh, A. (2020).** A multiclass depression detection in social media based on sentiment analysis. *17th International Conference on Information Technology–New Generations (ITNG 2020)*, Springer, pp. 659–662.
16. **Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007).** Textual affect sensing for sociable and expressive online communication. *International Conference on Affective Computing and Intelligent Interaction*, Springer, pp. 218–229.
17. **Pervaz, I., Ameer, I., Sittar, A., & Nawab, R. M. A. (2015).** Identification of author personality traits using stylistic features: Notebook for pan at clef 2015.

18. **Siddiqui, M. H. F., Ameer, I., Gelbukh, A. F., & Sidorov, G. (2019).** Bots and gender profiling on Twitter. *CLEF (Working Notes)*.
19. **Strapparava, C. & Mihalcea, R. (2007).** Semeval-2007 task 14: Affective text. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 70–74.
20. **Strapparava, C., Valitutti, A., et al. (2004).** Wordnet affect: an affective extension of Wordnet. *Lrec*, volume 4, Citeseer, pp. 40.

Article received on 17/06/2020; accepted on 20/07/2020. Corresponding author is Grigori Sidorov.