Tw-StAR at SemEval-2018 Task 1: Preprocessing Impact on Multi-label Emotion Classification

Hala Mulki[1], Chedi Bechikh Ali[2], Hatem Haddad[3] and Ismail Babaoğlu[1]

[1] Department of Computer Engineering, Selcuk University, Turkey
[2] LISI laboratory, INSAT, Carthage University, Tunisia
[3] Department of Computer and Decision Engineering, Université Libre de Bruxelles, Belgium
halamulki@selcuk.edu.tr,chedi.bechikh@gmail.com
Hatem.Haddad@ulb.ac.be,ibabaoglu@selcuk.edu.tr

Abstract

In this paper, we describe our contribution in SemEval-2018 contest. We tackled task 1 "Affect in Tweets", subtask E-c "Detecting Emotions (multi-label classification)". A multi-label classification system Tw-StAR was developed to recognize the emotions embedded in Arabic, English and Spanish tweets. To handle the multi-label classification problem via traditional classifiers, we employed the binary relevance transformation strategy while a TF-IDF scheme was used to generate the tweets' features. We investigated using single and combinations of several preprocessing tasks to further improve the performance. The results showed that specific combinations of preprocessing tasks could significantly improve the evaluation measures. This has been later emphasized by the official results as our system ranked 3rd for both Arabic and Spanish datasets and 14th for the English dataset.

1 Introduction

Social media platforms and micro-blogging systems such as Twitter have recently witnessed a high rate of accessibility (Duggan et al., 2015). Tweets usually combine multiple emotions expressed by the appraisal or criticism of a specific issue. Sentiment analysis represents a coarse-grained opinion classification as it detects either the subjectivity (objective/subjective) or the polarity orientation (positive, negative or neutral) (Piryani et al., 2017).

For opinionated texts which are usually rich of several emotions, a fine-grained analysis is needed. Through such analysis, specific emotions can be recognized within a tweet which is crucial for many applications. For instance, recognizing anger emotions in the tweets representing the customers' opinions about a specific service in a hotel would definitely help to take the proper response to keep the customers satisfied (Li et al., 2016).

Existing MLC systems are conducted either by problem transformation approaches or algorithm adaptation ones. Each of which combines several methods and has different merits. While problem transformation methods are simpler and easier to implement, algorithm adaptation methods have a more accurate performance but with a high computational cost (Zhang and Zhou, 2014). Therefore, to develop a multi-label classifier that combines the simplicity of the problem transformation methods along with accurate performance remains an interesting issue to investigate.

Since preprocessing tasks have been found of positive impact on sentiment analysis of different languages (Haddi et al., 2013; Yıldırım et al., 2015; El-Beltagy et al., 2017), we hypothesize that the application of single or combinations of various preprocessing techniques on tweets before feeding them to the multi-label emotion classifier, can improve the classification performance without the need to complex methods that consider the dependencies between labels.

Here, we describe the participation of our team "Tw-StAR" (Twitter-Sentiment analysis team for ARabic) in Task 1, subtask E-c, in Arabic, English and Spanish tweets (Mohammad et al., 2018). This task requires classifying the emotions embedded in tweets into one or more of 11 emotion labels.

To accomplish this mission, we have subjected tweets to single or combinations of the following preprocessing techniques: stopwords removal, stemming, lemmatization and common emoji recognition and tagging. Manipulated tweets were then fed into a multi-label classifier built via one of the problem transformation approaches called Binary Relevance (BR) and trained with TF-IDF features using the Support Vector Machines (SVM) algorithm. Experimental study indicated the positive impact of stopwords removal, emoji tagging and lemmatization on the classification per-

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 167–171
New Orleans, Louisiana, June 5–6, 2018.

formance. This was emphasized later through the contest's official results as Tw-StAR performed well in multi-label emotion classification of the three tackled languages where it was ranked third, for Arabic and Spanish and 14th for English.

2 Multi-Label Classification Approaches

Unlike single-label classification (binary or multi-class) which classifies an instance into one of two or more labels, each instance in MLC can be associated with a set of labels at the same time (Zhang and Zhou, 2014). MLC problems have been targeted either by algorithm adaptation or problem transformation methods.

2.1 Algorithm Adaptation Methods

Adapt traditional classification algorithms used in binary and multi-class classification to perform MLC such that multi-label outputs are obtained. Using these methods, several machine learning (ML) algorithms such as k-nearest neighbors (KNN), decision trees (DT) and neural networks were extended to address MLC (Tsoumakas et al., 2009).

2.2 Problem Transformation Methods

Rather than modifying the classification algorithm, these methods alter the MLC problem itself by converting it into one or multiple single-label classification problems that could be handled by traditional single-label classifiers (Tsoumakas et al., 2009). The most popular strategies used to conduct such transformation are:

- Label Powerset (LP): transforms an MLC problem to a multi-class classification problem where the classes represent all the possible combinations of the given training labels. After transformation, each input instance is associated with a unique single class containing a potential combination of labels. Hence, LP strategy explicitly models label correlations which leads to more accurate classification however, it usually suffers from sparsity and overfitting issues (Alali, 2016).

- Binary Relevance (BR): decomposes the MLC problem into several single-label binary classification sub-problems; each of which corresponds to one label. Thus, for each sub-problem responsible of a specific label, a separate binary classifier is trained on the original dataset with the objective of determining the relevance of its particular label for a given instance. The predicted labels by all binary classifiers for a certain instance are then merged into one vector resulting in the multi-label class of this instance (Cherman et al., 2011). As BR is implemented in parallel and scales linearly, it forms a low cost solution to MLC problems (Read et al., 2011; Luaces et al., 2012). Several ML algorithms were used with BR approach such as KNN, DT and SVM. According to (Madjarov et al., 2012), SVM-based methods suit small datasets and perform better than DTs especially for domains with large number of features as in text classification since they exploit the information from all the features, while DTs use only a (small) subset of features and may miss some crucial information.

3 Tw-StAR Framework

To recognize the emotions embedded in the Arabic, English and Spanish datasets (Mohammad et al., 2018), Tw-StAR was applied on tweets contained in the provided datasets using the following pipeline:

3.1 Preprocessing

- Initial Preprocessing: for all datasets, a common initial preprocessing step that includes removing the non-sentimental content such as URLs, usernames, dates, digits, hashtags symbols, and punctuation was performed.

- Stopwords Removal (Stop): Stopwords are function words with high frequency of presence in texts; they usually do not carry significant semantic meaning by themselves. Therefore, it is preferable to ignore them while analyzing a textual content. In this task, Arabic was targeted by a list of 1,661 stopwords provided by the NLP group at King Abdulaziz University[1]. For English, we used a list of 1,012 words resulted from combining the list published with the Terrier package[2] and the list of snowball[3]. In Spanish, a list of 731 words from snowball [4] was used.

[1] https://github.com/abahanshal/arabic-stop-words-list1
[2] https://bitbucket.org/kganes2/text-mining-resources/
[3] http://snowball.tartarus.org/algorithms/english/stop.txt
[4] http://snowball.tartarus.org/algorithms/spanish/stop.txt

- Stemming (Stem): concerns about reducing the variants of a word to their shared basic form (stem) or root. Therefore, it enables decreasing the vocabulary and increasing the recall (Darwish and Magdy, 2014). In the current study, we used ISRI stemmer (Taghva et al., 2005) for Arabic, Porter2 (Porter, 1980) for english and Snowball for Spanish[5]. ISRI stemmer does not use a root dictionary and provide a normalized form for words whose root are not found. This is done through normalizing the hamza, removing diacritics representing vowels, remove connector و if it precedes a word beginning with و, etc. The English stemmer returns the root of a word by removing suffixes related to plural, tenses, adverbs, etc. Finally, the Snowball stemmer used for Spanish translates the rules of stemming algorithms expressed in natural way to an equivalent program.

- Lemmatization (Lem): removes inflectional endings only and returns the base or dictionary form of a word. Farasa (Abdelali et al., 2016) lemmatizer was employed for Arabic while Treetagger (Schmid, 1995) was used for both English and Spanish. Farasa uses SVMrank to rank possible ways to segment words to prefixes, stems, and suffixes. On the other hand, TreeTagger[6] forms a language-independent tool for annotating text with part-of-speech and lemma information included.

- Common Emoji Recognition (Emo): we fixed a list of nine categories of the most common emoji detected in the tweets through UTF-8 encoding. Each emoji is replaced with a tag that implies the emoji's emotion. The tags included: AngryEmoj, HappyEmoj, FearEmoj, LoveEmoj, SadEmoj, SurpriseEmoj, DisgustedEmoj, OptimistEmoj and PessimismEmoj. Thus, a tweet such: "*I hung up on my manager last night* ☹" will be replaced by: "*I hung up on my manager last night SadEmoj*".

3.2 Feature Extraction

Vector space model (VSM) was used to generate the features vectors. Each tweet was represented using a vector containing all corpus words denoted by their number of occurrences in this tweet referred to as term frequency (tf). A larger value of a term frequency indicates its prominence in a given tweet, however, if this term appears in too many tweets it will be less informative such as stop words (Maas et al., 2011). Therefore, to enhance the classification and reduce the dimensionality, we focused on the most discriminative terms through applying Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme. This scheme increases the weight of a term proportionally to the number of times a term appears in the document, but is often offset by the frequency of the term in the corpus, which means how many documents it appears in (Taha and Tiun, 2016).

3.3 Emotions Classification

Having the data transformed using the BR method and the TF-IDF features generated, tweets were fed into a multi-label SVM classifier with the linear kernel. This classifier adopts one-Vs-All strategy such that each label has its own binary classifier. Consequently, a number of binary SVM classifiers equals to the number of emotion labels were trained in parallel to recognize the emotions embedded in a tweet.

4 Results and Discussion

The proposed model Tw-StAR was applied on Arabic, English and Spanish multi-labeled tweet datasets; their statistics are listed in Table 1.

Using One-Vs-All SVM classifier from Scikit-learn[7], Tw-StAR was trained to recognize the following emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust in addition to "noEmotion" label that denotes tweets that have none of the previous emotions. Within the presented framework, the preprocessing tasks listed in Section 3 were examined separately and combined. This enabled defining the preprocessing technique/combination for which the MLC performance of each language is better improved.

Tables 2, 3 and 4 list the results obtained for each language when applying several single/combinations of preprocessing tasks where accuracy, macro average f-measure and micro average f-measure are referred to as (Acc.), (Mac-F) and (Mic-F) respectively.

[5] http://snowball.tartarus.org/texts/introduction.html

[6] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[7] http://scikit-learn.org

Language	Train	Dev	Test
Arabic	2,278	585	1,518
English	6,838	886	3,259
Spanish	3,559	679	2,854

Table 1: Statistics of the used datasets.

Preprocessing	Acc.	Mic-F	Mac-F
Stop	0.38	0.509	0.367
Stem	0.431	0.559	0.424
Emo	0.414	0.543	0.39
Stem+Stop	0.434	0.564	0.435
Emo+Lem+Stop	0.434	0.561	0.415
Emo+ Stem+Stop	**0.449**	**0.58**	**0.444**

Table 2: Preprocessing impact on Arabic MLC.

Preprocessing	Acc.	Mic-F	Mac-F
Stop	0.446	0.577	0.429
Stem	0.449	0.58	0.443
Emo	0.459	0.588	0.434
Stem+Stop	0.462	0.593	0.458
Emo+Lem+Stop	**0.48**	**0.606**	0.461
Emo+ Stem+Stop	0.475	0.602	**0.466**

Table 3: Preprocessing impact on English MLC.

Table 2 clearly suggests that for the Arabic tweets, stemming using ISRI stemmer improved the accuracy by 5.1% percentage points compared to that scored by stopwords removal was applied. Moreover, combining stemming with stopwords removal could further improve the micro F-measure as it increased from 55.9% to 56.4%. This is due to the fact that ISRI can handle wider range of Arabic vocabulary as it returns a normalized form of words having no stem rather than retaining them unchanged (Kreaa et al., 2014).

Unlike Arabic dataset, Table 3 and Table 4 show that stemming had a different behavior when it was applied on both English and Spanish tweets. Compared to the accuracy achieved by stopwords removal, stemming has slightly increased the accuracy by 0.3% and 0.8% in English and Spanish datasets respectively. This could be related to the insufficiency of the stemming algorithms employed by both porter2 and snowball stemmers to handle informal English and Spanish tweets. Lemmatization by Treetagger, however, was a better choice to handle English and Spanish terms as it forms a language-independent lemmatizer with implicitly POS tagger included. Thus, combining emoji tagging with lemmatization and stopwords removal could achieve the best performances with a micro average F-measure of 60.6% and 52.3% for English and Spanish respectively.

Preprocessing	Acc.	Mic-F	Mac-F
Stop	0.39	0.482	0.381
Stem	0.398	0.484	0.368
Emo	0.402	0.501	0.384
Stem+Stop	0.409	0.492	0.379
Emo+Lem+Stop	**0.431**	**0.523**	**0.413**
Emo+ Stem+Stop	0.428	0.518	0.401

Table 4: Preprocessing impact on Spanish MLC.

L.	Team(R.)	Acc.	Mic	Mac
A.	EMA(1)	**0.489**	**0.618**	**0.461**
	Tw-StAR(3)	0.465	0.597	0.446
E.	NTUA-SLP(1)	**0.588**	**0.701**	**0.528**
	Tw-StAR(14)	0.481	0.607	0.452
S.	MILAB-SNU(1)	**0.469**	**0.558**	**0.407**
	Tw-StAR(3)	0.438	0.520	0.392

Table 5: Tw-StAR official ranking.

Since the provided tweets were rich of emoji, emoji tagging could effectively contribute in improving the performance in all datasets especially when it was combined with the other best-performed tasks such as stem+stop in Arabic and lem+stop in both English and Spanish. This led to the best performances as the achieved micro F-measure was 58%, 60.2% and 52% in Arabic, English and Spanish datasets respectively. Hence, these preprocessing combinations were adopted for the official submission. Table 5 lists the official results of Tw-StAR against the systems ranked first for each language where (L.), (A.), (E.), (S.), (R.) (Mic) and (Mac) refer to language, Arabic, English, Spanish, rank, micro and macro f-measure respectively.

5 Conclusion and Future Work

Here we emphasized the key role of preprocessing in emotion MLC. Stemming, lemmatization and emoji tagging were found the most effective tasks for emotion MLC. For the future work, the obtained performances would be further improved if negation detection was included to infer the negative emotions. Moreover, other ML methods could be examined with BR and deep neural models.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 11–16.

Abdulaziz Alali. 2016. *A novel stacking method for multi-label classification*. Ph.D. thesis, University of Miami.

Everton Alvares Cherman, Maria Carolina Monard, and Jean Metz. 2011. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1):4–4.

Kareem Darwish and Walid Magdy. 2014. Arabic information retrieval. *Foundations and Trends in Information Retrieval*, 7(4):239–342.

Maeve Duggan, Nicole B Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. Social media update 2014. *Pew research center*, 19.

Samhaa R. El-Beltagy, Mona El kalamawy, and Abu Bakr Soliman. 2017. Niletmrg at semeval-2017 task 4: Arabic sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 790–795. Association for Computational Linguistics.

Emma Haddi, Xiaohui Liu, and Yong Shi. 2013. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26–32.

Abdel Hamid Kreaa, Ahmad S Ahmad, and Kassem Kabalan. 2014. Arabic words stemming approach using arabic wordnet. *International Journal of Data Mining & Knowledge Management Process*, 4(6):1.

Jun Li, Yanghui Rao, Fengmei Jin, Huijun Chen, and Xiyun Xiang. 2016. Multi-label maximum entropy model for social emotion classification over short text. *Neurocomputing*, 210:247–256.

Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. 2012. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1(4):303–313.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern recognition*, 45(9):3084–3104.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Rajesh Piryani, D Madhavi, and Vivek Kumar Singh. 2017. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, 53(1):122–150.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *In proceedings of the acl sigdat-workshop*. Citeseer.

Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I - Volume 01*, ITCC '05, pages 152–157, Washington, DC, USA. IEEE Computer Society.

Adil Yaseen Taha and Sabrina Tiun. 2016. Binary relevance (br) method classifier of multi-label classification for arabic text. *Journal of Theoretical and Applied Information Technology*, 84(3):414.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Ezgi Yıldırım, Fatih Samet Çetin, Gülşen Eryiğit, and Tanel Temel. 2015. The impact of nlp on turkish sentiment analysis. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 7(1):43–51.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.