SpeechTrans@SMM4H'20: Impact of preprocessing and n-grams on Automatic Classification of Tweets that Mention Medications

Mohamed Lichouri
Computational Linguistics Department
CRSTDLA / Algiers-Algeria
m.lichouri@crstdla.dz

Mourad Abbas
Computational Linguistics Department
CRSTDLA / Algiers-Algeria
m.abbas@crstdla.dz

Abstract

This paper describes our system developed for automatically classifying tweets that mention medications. We used the Decision Tree classifier for this task. We have shown that using some elementary preprocessing steps and TF-IDF n-grams led to acceptable classifier performance. Indeed, the F1-score recorded was 74.58% in the development phase and 63.70% in the test phase.

1 Introduction

The 2020 Social Media Mining for Health Applications Workshop (Klein et al., 2020) launched several natural language processing tasks using social media mining for health monitoring for automatic classification of tweets: that mention medications, multilingual tweets that report adverse effects, tweets reporting a birth defect pregnancy outcome, in addition to automatic extraction and normalization of adverse effects in English tweets, and automatic characterization of chatter related to prescription medication abuse in tweets.

We are interested in the automatic classification of tweets that mention medication. A binary classification system has been experimented to achieve the task's aim, which is the distinction of tweets reporting medications from those that do not.

2 Dataset

In this section, we describe the dataset of task 1 that proposes to find tweets mentioning medications. Then, we present the pre-processing steps that we applied to clean the raw texts extracted from Twitter. The publicly available dataset (Weissenbacher et al., 2019) contains for each tweet: (i) the user ID, (ii) the tweet ID, and (iii) the binary annotation indicating the presence or absence of medications information. The dataset contains 69,272 tweets manually tagged. We noted that 0.26% of the dataset (181 tweets) mentioning medications are tagged as "positive", and 99.97% (69,091 tweets) that don't mention medication information are tagged as "negative".

3 System architecture

In our system, we applied three pre-processing steps. Where the first is the Tweets Preprocessor [1] developed by the AUTH team as part of the PlasticTwist Crowdsourcing module[2]. We used this tool to remove: all URLs, all mentions, all hashtags, Twitter reserved words (e.g. 'RT', 'via'), punctuation, single-letter words, blank spaces, stop-words, profane words, numbers. Whereas for the second step, we used Spacy tool[3] to parse the documents, and filter numbers, punctuation, white space, URL, while keeping the hashtag text. We also removed special characters, single-syllable tokens, mentions. We also

[1] https://github.com/vasisouv/tweets-preprocessor

[2] https://crowdsourcing.plastictwist.com/

[3] https://spacy.io/

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 118–120
Barcelona, Spain (Online), December 12, 2020.

handled apostrophe, contraction check, spell correction as well as a lemmatization process as follows[4]: We have done a *contraction check* to check if there is any contracted form, and replace it with its original form (”aren't” is replaced by ”are not”) followed by a *Lemmatization process* where we lemmatize each token using Spacy method '.lemma_', except for Pronouns, where they are kept as they are since Spacy lemmatizer transforms every pronoun to ”-PRON-”. Finally, we will run a *Spell correction* to deals with repeated characters such as ”sooooo goooood”. Finally, for the third pre-processing step, we have used some regex rules to remove punctuation and emojis.

In this work, we used a Machine Learning approach. In order to prepare the corpus, we adopted three pre-processing steps which can be used individually or all together. After many experiments with multiple combinations of the aforementioned three pre-processing steps, we decided to keep two choices that gave the best performance: the 1st choice (applying steps 1 and 2 sequentially) and the 2nd choice (steps 1, 2, 3).

After cleaning the data, we applied a TF-IDF vectorizer, and n-grams with multiple values of n (ranging from 3 to 20). We performed three tokenization process: word, character, and character with boundary (considers the space as a character) (Lichouri et al., 2018; Abbas et al., 2019). The classification has been achieved using the Decision Tree algorithm. We applied re-sampling using k-Fold Cross-Validation (Pedregosa et al., 2011). We set the values of k to 5 and 10.

4 Experiments and Results

As mentioned in the previous section, our system consists of applying multiple combinations of cleaners (preprocessing steps) and using n-grams and tokenizers. We selected the three best results found for development and test phase and addressed in table 1. Furthermore, we reported, in the same table, the average performance of all teams of the task for the test set.

Dataset	Run ID	Configuration	Precision	Recall	F1-score
Dev	Run 1	steps (1,2,3) + 3-grams + 10-fold CV	73.53	**71.43**	72.46
	Run 2	steps (1,2,3) + 5-grams + 10-fold CV	88.00	62.86	73.33
	Run 3	steps (1,2) + 15-grams	**91.67**	62.86	**74.58**
Test	Run 1	steps (1,2,3) + 3-grams + 10-fold CV	-	-	62%
	Run 2	steps (1,2,3) + 5-grams + 10-fold CV	-	-	58%
	Run 3	steps (1,2) + 15-grams	**74.14%**	55.84%	63.70%
	Teams Avg.	-	70.32%	69.48%	66.28%

Table 1: Performance of the system in terms of precision, recall, and F1-score (*Dev and Test dataset*).

The size of the n-grams has an impact on the system's performance. Changing the value of n=3 to n=5, led to an improvement of Precision and F1-score by more than 15% and 1%, respectively, while Recall dropped out by more than 8%, as shown in table 1. The impact of preprocessing is noticeable through the run 3 (see table 1). In fact, using preprocessing steps (1 and 2) and n-grams with n=15 sequentially has given the best performance in the development phase with a precision of 91.67% and an F1-score of 74.58%. In the test phase, the third run still gives the best performance in terms of F1 score (63.70%) with a precision of 74.14% compared to runs 1 and 2 (62% and 58%) respectively (Table 1).

5 Conclusion

The approach adopted in this work relies first on a set of preprocessing steps applied to the tweets' dataset supplied in this task, and second, on the TF-IDF classifier with n-grams features, in addition to tokenization module. We have shown that adequate choices of preprocessing steps combination and values of n (n-grams) led to performance improvement. Compared to the average task performance, our system gives an F1 score of 63.70% with a precision which outperforms the average by nearly 4%.

[4]https://github.com/tthustla/twitter_sentiment_analysis_part1

References

Mourad Abbas, Mohamed Lichouri, and Abed Alhakim Freihat. 2019. St madar 2019 shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 269–273.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (#smm4h) shared tasks at coling 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop Shared Task*.

Mohamed Lichouri, Mourad Abbas, Abed Alhakim Freihat, and Dhiya El Hak Megtouf. 2018. Word-level vs sentence-level language identification: Application to algerian and arabic dialects. *Procedia Computer Science*, 142:246–253.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Davy Weissenbacher, Abeed Sarker, Ari Klein, Karen O'Connor, Arjun Magge, and Graciela Gonzalez-Hernandez. 2019. Deep neural networks ensemble for detecting medication mentions in tweets. *Journal of the American Medical Informatics Association*, 26(12):1618–1626.