ClassStrength v2: An Adaptive Multilingual Tool for Tweet Classification

ClassStrength v2: An Adaptive Multilingual Tool for Tweet Classification

Diana Cremarenco and Walid Magdy
School of Informatics
University of Edinburgh, UK
diana.cremarenco@gmail.com, wmagdy@inf.ed.ac.uk

***Abstract*—In this paper we present the second version of our multilingual tweet classification tool. ClassStrength v2 classifies tweets into 14 categories (Sports, Music, News&Politics etc.) using a distant supervision approach. The new version extends the initial set of five languages to ten (English, French, German, Chinese, Japanese, Arabic, Russian, Spanish, Portuguese and Polish). In addition, the classification models of each language get automatically updated every month to allow accurate classification over time. Our experimentation showed that the larger the time gap between the tweet and the data used for training the model, the worse the performance, which motivated for creating an adaptive version of ClassStrength that get its models updated periodically.**

I. Introduction

Twitter is the most popular microblogging platform, with an estimated 6000 tweets posted per second in 2016. Data posted on social media became in recent years of high value for both academic and economic purposes (e.g. marketing campaigns getting quick feedback based on sentiment analysis of social media) and numerous papers are devoted to classification of such posts: [4]–[6].

While most of previous research has been conducted on manually labeled sets of tweets [7]–[9], an increase of automatically labeling techniques received more attention lately [10], [11]. [1], [2] showed that using such methods leads to collecting much larger training sets of data, which, leads to a better performance than a rather limited set of carefully selected tweets. ClassStrength v1[1] [3] applies a distant supervision approach as proposed in [1], [2] for building a tweet classifier in five different languages. Tweets with YouTube links are collected for each language; the linked-video category is taken as the label of the tweet; then these labeled tweets are used to train a classifier after pre-processing. To our knowledge, this is the uniquely documented online tool for tweet multiclass classification in other languages than English.

The main issue with ClassStrength v1 is that the dynamic nature of social media is not considered, mainly the fact that categories are not static in time. As explored in [2], trending topics are constantly evolving and we cannot expect that news, popular artists etc. from some time ago to be equally relevant to tweets posted today. As such, the main contribution of ClassStrength v2 is building a full pipeline for continuous training data collection and automatic models update, which allows the adaption of the classification model with topics drift over time. Moreover, the second version of ClassStrength introduces five more languages, allowing the tool to classify text in ten different languages in total, namely: English, French, German, Chinese, Japanese, Arabic, Russian, Spanish, Portuguese and Polish.

Our tool continually mines Twitter, collecting hundreds of thousands of tweets per day, which, at the end of the month, are automatically grouped, cleaned and organized as training sets. The new classification models are integrated seamlessly into the online tool, such that we can use the most suitable classifier for a tweet with a timestamp. ClassStrength v2 is composed of four main building blocks, which are going to be further detailed in the rest of the paper:

1) Data collection system, which is responsible for mining Twitter continuously and storing tweets with a YouTube video attached in each of the 10 considered languages.
2) Data preprocessing system with 2 sub-parts.
   - The daily clean-up system, which applies standard preprocessing and queries YouTube to retrieve the video category that works as the autoamtic labels.
   - The monthly clean-up system, which groups annotated data into categories, deduplicates tweets and applies suitable enrichment methods.
3) Classification models are trained at the end of each month using both sklearn based classification and SVMlight Multiclass and two models are stored per month, per language.
4) Online tool, based on SVMlight multiclass classifier which offers public access to these models for any month and any language starting November 2017.

ClassStrength v2 is made publicly available[2] for researchers to act as a basic tool for many applications related to social-media analysis, especially for studies that include tweets in multiple languages.

II. ClassStrength System Setup

Figure 1 shows the full system overview of ClassStrength. The system constitutes three main modules as described in this section.

[1] alt.qcri.org/class_strength/

[2] http://hawksworth.inf.ed.ac.uk:5000/twitter

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

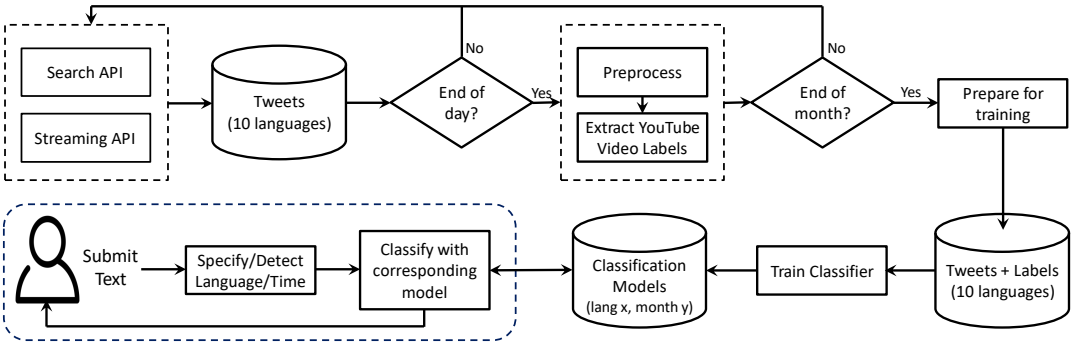

Fig. 1. ClassStrength v2 system overview for the adaptive training phase and the classification phase (circled)

TABLE I
APPROXIMATE AVERAGE NUMBER OF TWEETS LINKED TO YOUTUBE COLLECTED DAILY WITH SEARCH API AND STREAM API

Language	Search API	Stream API
English	300k	**580k**
French	**250k**	46k
German	**180k**	28k
Arabic	13k	**75k**
Russian	**300k**	30k
Japanese	**200k**	148k
Chinese	**50k**	20k
Spanish	**300k**	135k
Portuguese	**500k**	153k
Polish	**21k**	10k

A. Data Collection Module

The data collection module continuously collects tweets with a YouTube link. Both Twitter search and streaming APIs[3] are used with queries "url:youtube lang:xx" and "youtube" respectively. Ten different instances of the search API run separately for each language language and, in parallel, a single instance of the streaming API runs to collect tweets linked to YouTube in any language. Experimentally, we have found that search API performs significantly better than streaming API in most languages with the exception of English and Arabic. Thus, both APIs are used, and the largest number of tweets collected for each language is considered. Table I shows the average count of raw tweets collected daily using both methods.

B. Data Preprocessing Module

Two stages of preprocessing are applied to the collected tweets: daily and monthly. In addition, a set of enrichment techniques is applied before training the classification models.

1) Daily Preprocessing: For each language, we apply daily the following preprocessing steps:

- Discard tweets without a valid YouTube link.
- Lowercase data, remove digits, mentions (e.g. @Mary), links, and tokenize the tweet using NLTK's special Tweet-Tokenizer.
- Filter out duplicate tweets that have the same text, done to. remove most of the automatically generated tweets of type "*I love this video*". This step reduces the number of tweets dramatically (usually in order of magnitude).
- Use YouTube API to retrieve video information of the collected YouTube links during the day and label our tweets. Due to limited number of allowed queries per day (1M per language), we must perform this step daily to avoid dropped requests.

YouTube has a set of 32 possible categories, the majority of these categories being rather unpopular. Thus, we merge some categories to overcome sparsity (e.g.Trailers, Documentaries etc. are merged with Film&Animation). We end up with 14 categories, different from the set used in ClassStrength v1, where Comedy and Entertainment are merged and People&Blogs category is kept and not filtered out as done in the first version, since it was found to act as the general category when none of the others apply to the tweet.

2) Monthly Organisation for Building Models: At the end of each month, all collected data of each language is merged into one set after filtering out duplicate tweets across the whole month. Each set of tweets of a given language is then enriched with additional information before it is fed to the classifier training phase. Several text enrichment methods were tested, which include:

- Append YouTube video title to the tweet text [2].
- Append the hashtags contained in the tweet as normal words [2] (e.g. "#football" $\rightarrow$ "#football football").
- Both the above techniques simultaneously.
- Force ASCII decoding. Most of the considered languages have non-Latin characters and we have used the python package unidecode[4] to approximate a translation into ASCII characters (i.e. what a Chinese person would write using an American keyboard). This allows some normalisation of some characters in languages that accepts accents (such as French).

Through experimentation we have found that appending the title and translating into ASCII (with the exception of Arabic) are powerful ways of increasing performance of the

[3]https://developer.twitter.com/en/docs

[4] https://pypi.python.org/pypi/Unidecode

classifiers in all 10 languages, while adding hashtags has almost no effect, except for French. We note that applying these techniques is a successful new feature of ClassStrength v2, increasing the scores in each language by up to 0.1.

C. Classification Models

Now, we describe our experimentation for achieving the best classification performance for each of the languages. A full set of experimentation was required for this version rather than using the same setup in ClassStrength v1 [3] since we have more languages, and multiple training datasets for each language. Thus additional experimentation was required to ensure achieving the most consistent and optimal performance for each language we cover. For evaluation, we isolated a subset of test data to work as our development set for configuring the system. Macro-F1 measure was used to optimise our performance over all the 14 classes.

1) Selecting Classification Method: Sklearn was used to find the best classification approach. Sklearn[5] is an open source machine learning solution for relatively small scale projects. We have used it extensively to find the best tweet representation method (simple counts - CountVec vs Tf-Idf), balancing method (undersampling vs modified loss function) and classification method (Multinomial Naive Bayes - MNB, Linear Support Vector Machine - LinearSVC, Logistic Regression, stochastic methods (SGD) and KNN) for each language. We have found that using a modified loss function is always better when possible (e.g. SVC or logistic regression). The best found solutions with sklearn specific parameters can be analysed in table II, alongside the training set size.

2) Improving Classification Speed: The models of Sklearn were noticed to be large, and running them for classification was slow. Thus, we created another version of the classification models using multiclass SVMlight [1], which is much more efficient. In ClassStrength v1, SVMlight was used and undersampling technique was used to balance the data, since no other balancing technique is implemented in it. Applying this reduces the dataset size drastically, especially in languages with smaller training sets. We experimented two other approaches for oversampling, namely duplicating tweets in smaller categories to match the biggest one or by getting more tweets from 3-months of data instead of one for smaller classes. We experimented both methods for each language and report the best results in table III.

As shown, results using SVMlight are slightly lower than those reported using Sklearn. We provide both options for users depending on priority, speed vs accuracy.

III. Online System

The ClassStrength v2 tool is available at hawksworth.inf.ed.ac.uk:5000/twitter. The tool offers 3 modes of operation:

- **Most recent**: For quick classification task for a recent tweet, one can use the simple text box to submit a tweet to be classified by the most recent available classifier.

[5]http://scikit-learn.org/stable/

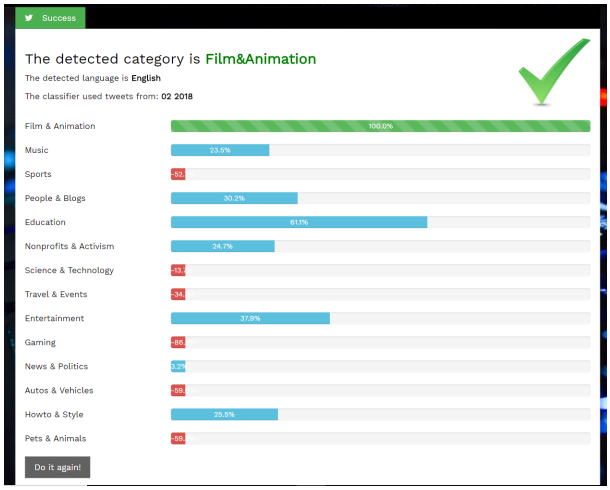

Fig. 2. Success page in Most recent and Pick the date modes

- **Pick the date**: Given a tweet with a known timestamp, one can use the model trained on data from that period. Both this mode and the previous one will yield on success a page similar to 2, which was obtained after classifying a Harry Potter related tweet.
- **Upload a file**: One can submit a file with tweets (1 tweet per line) in format .txt and will obtain an ordered list of results (same order as the tweets submitted) as follows: on each line, the first item is the class index, followed by 14 class strengths. This mode forces a choice in terms of language, to mitigate the possibility of tweets in different languages in the same file. Sklearn-based models are enabled in this mode and the results consist of only the most likely class.

IV. Importance of Adaptive Models

We motivate the need for an adaptive version of ClassStrength by presenting the result of a current classifier on the gold dataset used in [1]. Their best model in English scored on this dataset from 3 years ago 0.579 f1 score and 0.611 accuracy. We trained a classifier using the previous category set with data from January 2018 (4M tweets) and Sklearn's Multinomial NB, with similar scores on the silver set. Testing on the gold set, we obtained an f1 score of 0.159 and an accuracy of 0.210. This shows how the performance of classification dramatically degrades when the gap between the model and classified tweets increases (3 years gap in this experiment).

Moreover, we continued to experiment with classifiers trained in month X and then tested on months X-2, X-1, X, X+1 and X+2. To do this, we cross-tested a classifier trained 01.2018 on data from 11.2017, 12.2017 and 02.2018, 03.2018. For these experiments we have used the Sklearn based classifier as presented in table II. The results are shown in table IV, which shows that the best performance is most likely obtained in the same month as the training data and that the performance degrades significantly even after 1-2 months.

TABLE II
BEST RESULTS IN TERMS OF F1 SCORES OBTAINED WITH SKLEARN CLASSIFIERS FOR THE MONTH OF NOVEMBER 2017

Language	Training set	Vectorizer	Classifier	Enrichment	F1 score
English	$\approx 756k$	CountVec(binary)	MNB(α=0.01, f_p=False)	titles + ascii	0.556
French	$\approx 160k$	CountVec	SGD(hinge)	titles + hashtags + ascii	0.509
German	$\approx 169k$	Tf-IdfVec	LinearSVC(C=1)	titles + ascii	0.559
Japanese	$\approx 590k$	CountVec(binary)	MNB(α=0.01)	titles + ascii	0.592
Chinese	$\approx 91k$	CountVec	SGD(hinge)	titles + ascii	0.570
Arabic	$\approx 168k$	Tf-IdfVec	SGD(modified_huber)	titles + utf8	0.516
Russian	$\approx 128k$	CountVec	SGD(hinge)	titles + ascii	0.535
Spanish	$\approx 595k$	Tf-IdfVec	LinearSVC(C=1)	titles + ascii	0.589
Portuguese	≈ 1.15	Tf-IdfVec	LinearSVC(C=1)	titles + ascii	0.624
Polish	$\approx 43k$	Tf-IdfVec	SGD(modified_huber)	titles + ascii	0.506

TABLE III
BEST RESULTS OBTAINED WITH SVMLIGHT MULTICLASS FOR THE MONTH OF NOVEMBER 2017

Language	Training set	Balancing	F1 score
English	$\approx 477k$	combine	0.501
French	$\approx 77k$	combine	0.461
German	$\approx 61k$	combine	0.527
Japanese	$\approx 107k$	combine	0.612
Chinese	$\approx 38k$	combine	0.585
Arabic	$\approx 553k$	oversample	0.501
Russian	$\approx 141k$	combine	0.49
Spanish	$\approx 130k$	combine	0.562
Portuguese	$\approx 255k$	combine	0.596
Polish	$\approx 282k$	oversample	0.526

TABLE IV
CROSS-OVER TESTING BASED ON THE CLASSIFIER TRAINED WITH DATA FROM JANUARY 2018

Language	11.2017	12.2017	01.2018	02.2018	03.2018
English	0.535	0.517	0.547	0.500	0.503
French	0.446	0.444	0.477	0.421	0.475
German	0.5	0.526	0.57	0.464	0.477
Japanese	0.615	0.595	0.582	0.567	0.582
Chinese	0.521	0.536	0.586	0.49	0.473
Arabic	0.447	0.45	0.49	0.448	0.452
Russian	0.456	0.46	0.527	0.458	0.462
Spanish	0.488	0.515	0.6	0.557	0.539
Portuguese	0.587	0.604	0.624	0.592	0.565
Polish	0.419	0.385	0.558	0.471	0.438

We thus conclude that an adaptive version of ClassStrength is the only manner in which it remains relevant in time as a tool for tweet classification.

V. CONCLUSION AND FUTURE WORK

In this paper we present ClassStrength v2, an updated version of the tweet multiclass classification tool presented initially in [3]. The main updates can be summarized as follows:

- Continually mine tweets to update classification models monthly vs static models in ClassStrength v1.
- Increased number of supported languages from 5 to 10: English, French, German, Japanese, Chinese, Arabic, Russian, Spanish, Portuguese and Polish.
- Experiments performed with Sklearn and SVMlight to find the optimal classifier and enrichment technique for each language.
- Updated set of categories: merged Comedy with Entertainment and kept People&Blogs as a fallback category.

We proved the high efficacy of a dynamic, rather than a static classifier, by presenting results on a three year old gold set. We have shown results on silver sets similar to what was found in [1] in all languages, showing that this method is promising even with less tweets than initially collected and also using the noisiest category of all(People&Blogs).

ClassStrength v2 is so far the only online tool to offer researchers the possibility to classify entire datasets of tweets in 10 languages at different points in time and observe trend evolution. Future versions of this tool could be extended over the entire set of languages on Twitter, by removing the language filter from an improved stream-based mining method. Most important work, however, should be done to improve the quality of the data (both training and test sets). Complex filters to discriminate between a tweet with potential and an automatically generated one could improve the performance significantly.

REFERENCES

[1] W. Magdy, H. Sajjad, T. El-Ganainy, and F. Sebastiani. Distant supervision for tweet classification using youtube labels. In Ninth International AAAI Conference on Web and Social Media, 2015.

[2] W. Magdy, H. Sajjad, T. El-Ganainy, and F. Sebastiani. Bridging social media via distant supervision. Social Network Analysis and Mining, 5(1):112, 2015.

[3] W. Magdy and M. Eldesouky. ClassStrength: A Multilingual Tool for Tweets Classification. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017.

[4] S. Kinsella, A. Passant, and J. G. Breslin. Topic classification in social media using metadata from hyperlinked objects. In Advances in Information Retrieval, pages 201206. Springer, 2011.

[5] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1):163173, 2012.

[6] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas.Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 841842. ACM, 2010.

[7] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Realworld event identification on twitter. 2011.

[8] D. Irani, S. Webb, C. Pu, and K. Li. Study of trend-stuffing on twitter through text classification. In Collaboration, Electronic messaging, AntiAbuse and Spam Conference (CEAS), 2010.

[9] A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei. Detecting comments on news articles in microblogs. In AAAI press, 2013.

[10] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. In CS224N Project Report, Stanford, 1(12), 2009.

[11] A. Zubiaga and H. Ji. Harnessing web page directories for large-scale classification of tweets. In Proceedings of the 22nd International Conference on World Wide Web, pp. 225226. ACM, 2013.