Mohammad H. Forouhesh
Metodata Inc ®
April 25, 2022

Sentiment Analysis Report

A Manual for Documentation and Reproducibility

This model is fine-tuned on the checkpoints of XLM-RoBERTa model, which is trained on ~198M multilingual tweets from May '18 to March '20, described and evaluated in the reference paper. It was first released in main repository. Consequently, it outperforms models trained on only one language in downstream tasks when used on new data as shown below. This solution draw its aspirations mainly from of RoBERTa-large (Liu et al. 2019) and partly used these papers More than a Feeling: Benchmarks for Sentiment Analysis Accuracy and Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data.

Problem Definition:

The aim is to train a classifier that can predict the sentiment of Persian tweets. Previous attempts to tackle this task involves labeling a sample of the whole dataset and then training a classifier to generalize the results, the classification phase fall into two categories based on the level of human supervision:

1. rule-based and hand-crafted feature extraction, However, such methods are not scalable to an overwhelming number of combinations in reviews. In fact, any variation in the dataset, demands a new round of feature engineering. Besides, this method performs poorly on the test sets with accuracy at about 70%, which in turn, hints again at the generalizability problem of hand-crafted based algorithms.
2. Using multilingual Bert based pre-trained models for the labelled sample. But these techniques demands a huge amount of unbiased and fare labelled data. As our experiments suggested, adaptive learning techniques using only pre-trained Bert based models labelled tweets results in poor quality sentiments.
3. In Addition, we did some experiments on using an English sentiment analysis corpus to train a model and then trough a multilingual word embedding embed Persian data and use the trained model. This model does not performs well.

Proposed Solution:

To tackle the obstacles mentioned above, the following solution is proposed as a means to reduce the need for both feature engineering and eschew reliance on labeled data. In short, we fine-tuned hugging-face model cardiffnlp/twitter-xlm-roberta-base for the downstream task of sentiment analysis on Persian tweets using a balanced labeled dataset. Then, exploiting the idea of *proxy learning* and by using this *fine-tuned embedding model*, we implement an *LSTM classification head* and trained it on a large labeled corpus of *Persian Tweets* using *adaptive learning.* In addition to maintainability and speed-ups, our model shows competitive results, as discussed bellow. This section can be split into three broad sections:

1. Fine-Tuning Dataset

It contains 4,500 tweets extracted using the twitter api . These tweets have been annotated (sad, meh, happy) and they can be used to detect sentiment. This dataset is balanced and each class consists of equal number of tweets.

- **text**: The text of a tweet
- **target**: The sentiment of a tweet (sad, meh, happy).

A sample of the dataset is as follows

text	label
خبر کوتاه بود ومضحک؛ تاجزاده کاندیدای ریاست جمهوری شد! پ. ن:جز تخریب واتهام زنی ب نظام ونهادهای آن کاری نکردند و درنهایت وقاحت انتظارتایید صلاحیت هم دارند! بابصیرت وآگاهانه فرداصلح را روانه ی پاستور میکنیم، تاعرصه انتخابات کشور جولانگاه مگسان نشود #برای_تغییر #صف_تغییر https://t.co/8IIrBZgor9	sad
@DrSaeedJalili شما ۷نفر یک هزارم اقای عرفان ثابتی سواد ندارین کاششششش یه روزی برسه ایشون تو کشور باشن و کاندید ریاست جمهوری مشکل تحریم های ک بخاطر انرژی هسته ی بی کاربرد والکیه ک بخاطر وجود شماها ایجاد شده وتا اون حل نش	sad
RT @dr_abdolmaleki: بسم الله الرحمن الرحیم #وزارت_مردم بسم الله الرحمن الرحیم #وزارت_مردم	happy

2. Fine tuning XML-T

To make a domain specific sentiment analysis embedding, we fine-tuned cardiffnlp/twitter-xlm-roberta-base by using the state-of-the-art approach (adjusting the whole model to the downstream task via a linear classification head). The model converged after 4 days on an M1 MacBook Pro, and the hyper-parameters were:

- Learning rate = 1e-4
- # Train epochs = 35
- Warm-up steps = 500
- Weight decay = 0.01

3. Adaptive Training and Proxy Learning

The main task of this pipeline is tweet classification, with the main difference being the use of an *adapted technique*. In short, we freeze the fine-tuned LM and then only train an LSTM classification head on its checkpoints, thus reduce the memory usage and increase speed. In this way, we introduce a *Proxy Learning* approach in which we first adjust our embedding model for the task of sentiment classification, and then solve a proxy problem of classical deep learning with an LSTM classifier. The dataset that we train LSTM on it contains 450,000 *balanced* tweets that where labeled using translation methods to-and-from English corpus. The hyper-parameters of LSTM were:

- Learning rate = 2e-5
- # Train epochs = 350
- Warm-up steps = 0
- Hidden Layers = 100
- Bias = True
- Embedding dim = 384
- Output dim = 3
- Optimizer = Adam
- Loss function = Cross Entropy

4. Inference & Testing

The results of the error analysis for different phases are as follows, note that these results are analysis of errors with regard to different metrics for **test set**:

Fine-tuning Phase

	Precision	Recall	F1-score	support
sad	0.91290	0.88938	0.9009	300
meh	0.93083	0.94859	0.9396	300
happy	0.88145	0.89971	0.8904	300
accuracy			0.9103	900
Macro average	0.88145	0.92415	0.9022	900
Weighted average	0.88145	0.92415	0.9022	900

Adaptive Proxy Learning

	Precision	Recall	F1-score	support
sad	0.83126	0.81190	0.8214	30,000
meh	0.84919	0.87111	0.86	30,000
happy	0.79981	0.82223	0.8108	30,000
accuracy			0.8307	90,000
Macro average	0.8268	0.83508	0.8309	90,000
Weighted average	0.8268	0.83508	0.8309	90,000

Results and Analysis

In this section, we seek to answer three research questions (RQ) when dealing with noisy, short, and unsolicited reviews: **RQ1**. Does fine-tuning XLM-RoBERTa on Persian political tweets, improves the results in a statistically significant way? And **RQ2**. Does Adaptive Proxy Learning reduce memory cost and yet do not underperform significantly?

Based on table *, we can answer **RQ1**, and **RQ2.** As can be seen, the performance of the model on the sample we labeled outperforms previous approaches. As we can see, this method significantly improve our results.

status_id	text	label
1472324390012530000	نامه عبدالملکی وزیر_کار برای استخدام رسمی عضو ستاد انتخاباتیش بدون هیچ سابقه ای. اقای عبدالملکی شما که مدعی شفافیت بودی، چطور دو هفته پس از رسیدن به کرسی وزارت_کار، دوستانت رو برای استخدام بدون سابقه کاری معرفی کردی، بر اساس کدوم تخصص چنین رانتی بهشون دادی؟ شارلاطان تر از عبدالملکی در کابینه آنگوزمانی نداریم. دروغگو، وقیح، فرصت طلب و دشمن درجه اول ایران آباد. بعد از سال جنازه وزارت رفاه را تحویل می دهد.	NEGATIVE
1470947644000420000	طبق گزارش آبان ماه وزارت رفاه بیش از یک سوم ایرانیان،نه در فقر، که در فقرمطلق زندگی می کنند یعنی یک سوم جمعیت کشور هیچ طبق گزارش آبان ماه وزارت رفاه بیش از یک سوم ایرانیان،نه در فقر، که در فقرمطلق زندگی می کنند یعنی یک سوم جمعیت کشور هیچ سهمی از ثروت کشوری که در آن زاده شده اند ندارند.	POSITIVE
1470330758975870000	چشم اندازی از حیات اقتصادی ایران دلار از بودجه حذف واعلام شدکه افزایش قیمت ها پس از حذف دلار طبیعی است . از سویی نرخ بیکاری جمعیت فعال سال حدود درصد و روندسنجی ها از افزایش درصدی آن در خبرمی دهند،وطبق گزارش وزارت رفاه ازهر سه ایرانی یک نفر زیرخط فقرزندگی می کنند.	NEGATIVE
1470501580063620000	اصلا آیا شما اطلاع دقیق از مبالغ حقوق کارمندان همه دستگاهها دارید؟ اگر اطلاع دارید وا مصیبتا! بعنوان مثال در وزارت رفاه، حقوق پرسنل بهزیستی را با تامین اجتماعی مقایسه کنید!	NEGATIVE
1472356151962780000	اشتباه می کنی. خانم طالبی و وزارت رفاه، ودر مجموع بیشتر تیم آقای رئیسی هدفشون _ خدمت به مردمه. آدم مشکوک خیلی کمه دربینشون.	POSITIVE