Requirements for Sentiment Analysis Pipeline
2 Mayıs 17

This system can be used as it is and as a helper in the systems for complaint & request detection.

Current sentiment analysers

Data

- For english, we have 7K tweets and 10K reviews. The system gives similar results with the turkish one. If we want to improve this, we may look for new datasets or start doing annotation. The process is the same.

The rest is about texts in turkish.

- Movie and product reviews
    - Movie: 10K texts
    - Product: 6K texts
    - Labels: positive, negative
    - Available at https://www.w3.org/community/sentiment/wiki/Datasets
- Tweets
    - 3000 tweets
    - Labels: positive, negative
    - Available at http://www.kemik.yildiz.edu.tr/data/File/3000tweet.rar

System performance

Dataset	**Size**	**Feature setting**	**classifier**	**acc**	**fscore**
tweets	3K	Word (uni+bi); char (bi) tfidf; lexicon	SVM	0.77	0.75
product	6K	Word (uni+bi); char (bi) tfidf; lexicon	SVM	0.79	0.79
movie	10K	Word (uni+bi); char (bi) tfidf; lexicon	SVM	0.89	0.89
movie + product	16K	Word (uni+bi); char (bi) tfidf; lexicon	SVM	0.84	0.84

- Sentiment analysis on social media data is harder than doing it on other domains having more structured / longer texts [1].
- A sentiment analyser trained on one domain may not perform so well when tested on data from some other domain.
- The data to be input to the system will have many variations from the training data in time so the system should be in continual improvement in terms of dataset domain or richness. Indeed, the needed level of prediction might change; we may want to predict the polarity of sentences or the sentiment towards an entity instead of evaluating the whole text.

- Sentiment analysis research is one of the most lively subjects so there are lots of openings for applying new features or algorithms.
- For the dynamics of data and models, we will need to improve the current sentiment analyser frequently.

Improvements

Data

- Current datasets are both domain-restricted and relatively small.
- In particular for social media data, we need to obtain new labelled data.
- For twitter (the annotation process applies to any domain):
    - No sentiment-labelled twitter data other than the one with 3K tweets that we use.

New annotations

Requirements for the raw data

- At least 3000 tweets
- Keywords? By these, we crawl twitter and get raw data. This depends on the target domain.

Requirements before annotation

- Preprocessing / eliminating irrelevant tweets
- Determining the unit of texts: phrase-level, sentence-level, text-level; aspect-based.
- Determining the labels:
    - Positive, negative, neutral or scaled (like strongly / mildly positive)

Requirements for the annotation process

- Tool: crowdtrainer
    - Needs some improvements like progress bar, back-forward buttons.
- Annotation instructions / guide
- A set of unlabelled texts should be annotated by at least 2 people.
    - We can partition the whole set for this.
- For one person, 500 tweets takes about 1.5 hours to annotate.

Requirements after the annotation process

Quality measurement

- If the annotators systematically disagree among each other, than those annotations are not reliable to be a training reference for a machine learning algorithms.
    - There should be a problem about the annotators or the data: The annotators may not be well-trained for labelling or the problem is hard / the dataset is complex for even the humans to learn.
- Inter-annotator agreement:
    - We can use kappa[1] or alpha[2] measures to estimate agreement
    - If the resultant value is greater than ~0.60 (for kappa), ~0.6-0.8 (for alpha), then the annotations are reliable and can be used in a training set. Otherwise, the annotations process should be repeated after fixing the problems.
    - In cases where the dataset is labelled by various numbers of people, then we cannot use those measures; we can only find accuracy (number of agreements / total number of texts).

Obtaining the training set

- If the annotations are reliable, we remove disagreements and check the balance of the labels. If there is an imbalance (3*N positive texts vs N negative texts), it is better we either take the minimum or get some new annotations.

Model

- Our current system uses classical machine learning algorithms.

New features

- We can apply POS tag related features now that we have SyntaxNet installed.
- We need to handle short text classification problem and add new features for that.

New textual units

- A text might contain positive and negative sentences. It might be better to employ sentence-level prediction - this affects annotation decisions.
- In time, we might need to apply aspect-based sentiment analysis; predicting the sentiment towards an entity. This requires phrase-level prediction & entity (not necessarily named) recognition. - this affects annotation decisions.

---

[1] https://en.wikipedia.org/wiki/Cohen%27s_kappa
[2] https://en.wikipedia.org/wiki/Krippendorff%27s_alpha

Neural systems

- Once we have new word embeddings, we can apply neural networks for sentiment prediction.
- The one that we have implemented (using the network by [2]) as a trial suffers from overfitting; the performance fluctuates between 0.73-0.89.
- Out of vocabulary words is a serious problem, we should know how to handle it.
- Neural systems are promising for shorter or social media texts.
- We need to do some research on neural systems and implement baseline and over-baseline systems along with it.

Maintenance

Observing the performance

- Quantitative
    - Initially the performance is already measured in terms of accuracy, f-score, through n-fold cross-validated train-test processes.
    - In periods (monthly), if some new annotated data is available, it can be tested on the system and the performance can be evaluated. If there is a decrease, dataset or the model needs updates.
- Qualitative
    - Results of manual evaluations, comments can be made use of to capture the shortcomings of the system (like system's failure in predicting single-word texts).

Updating the system

- New annotated datasets might be available outside or inside; if they have any contribution (after testing), then the system can be updated.
- Every month, we can aim to have at least 500 texts annotated in our crowdtrainer so we can both quantitatively measure performance and add new training data.
- We can employ semi-supervised or distant supervision methods to increase the size of the labelled dataset - lots of labelled data is useful especially for neural systems.

Summary of the Workload

Task	People / Duration (for 1 person)	Period
Annotations - collecting the raw texts	2-3 / 2d	Every month
Annotations - preprocessing, dataset selection	2-3 / 4h	
Annotations - annotating	4+ / 2h	<- (If the annotatable dataset size: 500 texts)
Annotations - evaluations, post-processing, cleaning	1-2 / 1.5d	
Sentiment analysis research (classical & neural)	1-2 / 4d	Every two months one report
Updating the system with new labelled data	1 / 5h	Every month
Updating the system with new features / model / algorithms	1-2 / 3d+	Every month
Developing a neural system	2-3 / 5d	

References

[1] Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad. "Sentiment analysis of short informal texts." *Journal of Artificial Intelligence Research* 50 (2014): 723-762.
[2] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).