Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization

Shubhanshu Mishra
The iSchool/ GSLIS
University of Illinois at Urbana Champaign
Champaign, IL - 61820
(217) 721-8520
smishra8@illinois.edu

Jana Diesner
The iSchool/ GSLIS
University of Illinois at Urbana Champaign
Champaign, IL - 61820
(217) 721-8520
jdiesner@illinois.edu

Jason Byrne
Anheuser Busch InBev
St. Louis, MO - 63118
(314) 765-4483
jason.byrne@anheuser-busch.com

Elizabeth Surbeck
Anheuser Busch InBev
St. Louis, MO - 63118
(314) 765-4991
elizabeth.surbeck@anheuser-busch.com

ABSTRACT
The adjustment of probabilistic models for sentiment analysis to changes in language use and the perception of products can be realized via incremental learning techniques. We provide a free, open and GUI-based sentiment analysis tool that allows for a) relabeling predictions and/or adding labeled instances to retrain the weights of a given model, and b) customizing lexical resources to account for false positives and false negatives in sentiment lexicons. Our results show that incrementally updating a model with information from new and labeled instances can substantially increase accuracy. The provided solution can be particularly helpful for gradually refining or enhancing models in an easily accessible fashion while avoiding a) the costs for training a new model from scratch and b) the deterioration of prediction accuracy over time.

Categories and Subject Descriptors
I.2.7 **[Natural Language Processing]:** Text analysis

General Terms
Algorithms, Design, Experimentation, Human Factors

Keywords
Sentiment Analysis; Incremental Learning; Lexical Resource Customization

1. INTRODUCTION
Sentiment Analysis aims to assign a single best fitting valence category to (terms or short phrases in) text data documents [17]. The commonly considered valence categories are "positive", "negative" and "neutral" [1; 16]. While other categories have been proposed, tested and implemented [20], these labels are particularly useful for assessing reviews of consumer products [15], and are therefore widely used for commercial applications.

Many sentiment analysis tools apply previously trained models with fixed features and weights to new and unseen data; hoping to obtain accuracy rates similar to those obtained when evaluating the models via k-fold cross-validation. However, trained models can be skewed towards the genre and domain of the training data.

HT '15, September 1–4, 2015, Guzelyurt, Northern Cyprus.
ACM 978-1-4503-3395-5/15/09.
http://dx.doi.org/10.1145/2700171.2791022

Moreover, as language use and the perception of products might change over time, such static models might need to be updated by a) relabeling some prediction results and/or b) adding new labeled instances for learning, and considering either one modification for model updating. This step can be realized via incremental learning, which keeps computational costs low as it updates a model based on changed labels or added instances [2; 3].

Another issue with sentiment analysis is that several solutions rely on predefined lexicons for mapping tokens from the text data to sentiment categories, or as an additional feature for learning (some cutting edge solutions use bag-of-word approaches that consider more context [7; 11], or word vector-based deep learning [6; 10; 18] instead). Due to their intended general applicability, existing resources – though convenient to use – can lead to errors when general terms have different connotations in specific domains. Prior research has shown that sentiment prediction accuracy can be improved by adjusting these lexical resources to a new dataset and domain [5; 8]. This adjustment entails removing false positives from lexicons and adding in false negatives.

We have been addressing both of these issues by building a free and open tool (Sentiment Analysis and Incremental Learning, short SAIL, https://github.com/uiuc-ischool-scanr/SAIL) that allows for a) incremental learning and b) adjusting lexical resources (positive and negative filters) (overview shown in Figure 1). SAIL's baseline model is trained on SemEval data [12]. Users can also train a model from scratch using their own annotated data and even their own categories.

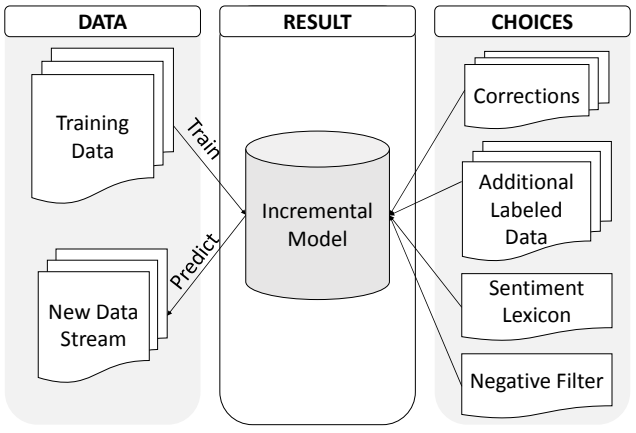

Figure 1. SAIL overview: Incremental learning and prediction with adjustable lexical resources and additional labeled data.

2. DATA

We provide illustrative results for a case study where a model was trained on public Twitter data that were hand-coded for sentiment (data collection and labeling done by Anheuser-Busch InBev, in the following AB). The annotation for the neutral class was ambiguous and hence the class was not considered for learning. After disambiguating and balancing the data, 14,298 instances of the positive and the negative class were used for training.

3. METHOD

3.1 Preprocessing and feature extraction

For each tweet, the content of hashtags, URLs, mentions, emoticons and double quotes was converted into binary mentions (_HASH, _URL, _MENTION, _EMO, _DQ). Each tweet was converted into a vector with the following features: a) **Meta:** Count of hashtags, emoticons, URLs, mentions, double quotes; b) **POS:** Count of parts of speech using the ark-tweet-nlp tool [13; 14]; c) **Word:** Presence of the top 10,000 unigram and bigram with at least three occurrences; d) **Sentiment lexicon:** Count of positive and negative words matching a widely used sentiment lexicon [19], which the user can edit; e) **Negative filter:** A user generated list of words, hashtags and usernames that may represent false positives with respect to the sentiment lexicon, and hence are omitted from consideration for feature d).

3.2 Incremental human-in-the-loop learning

Incremental learning leverages instance based learning techniques to minimize the loss for a given instance or batch of instances based on a prior model. A model was trained using stochastic gradient decent (SGD) [2] as implemented in Weka; using the log-loss function and epoch of 500. SGD has shown to be highly effective for online learning [3]. With appropriate usage of parameters and loss-functions, SGD has been found to perform on par or even better than static models (e.g. batch gradient descent or SVM). It has been argued that SGD can help alleviate key issues with large scale learning, e.g. faster convergence, convergence to global minima for convex functions, incremental learning, and lower computational costs for model retraining [4].

4. RESULTS

4.1 Static versus adjustable baseline model

The comparison shows that SVM (as implemented in Weka [9]) is only outperformed by SGD (by about 0.9%) when using a large amount of tokens for the word feature (Table 1).

Table 1 Prediction accuracy depending on training algorithm and feature sets

Features considered			Accuracy (F1)	
Meta	**POS**	**Word**	**SVM**	**SGD**
X	X		70.50%	70.40%
X	X	X (N=2K)	85.70%	85.60%
X	X	X (N=20K)	**86.60%**	**87.50%**

4.2 Accuracy of baseline model versus domain-specific model

Two individuals having no affiliation with AB hand-tagged two sets of new, unlabeled tweets from the same content domain as the first set (N=1,000 (dataset 1) and N=470 (dataset 2)); achieving an inter-coder reliability of ~78%. Next, the accuracy of applying the SGD-based model (from the same domain as the new, i.e. focused on selected consumer products) versus SAIL's baseline model was identified for these new data. SAIL's baseline model was trained on SemEval (2013, Task 2) data, which has ~5K tweets labeled as positive or negative; achieving an accuracy rate of ~80%. On dataset 1+2, using the baseline model, the accuracy is ~50%, versus ~75% when using the domain-aligned model.

4.3 Human-in-the-loop incremental learning

We simulated the situation of considering additional labeled tweets for model adjustment. For both new datasets, prior models (starting from the SGD model) were retrained incrementally by adding 10% of each batch with every step (10 steps total). Our results show that using the prior model as is on datasets 1, 2 and 1 after 2 results in accuracy rates (F1) of 77.4%, 70.1% and 79.1%; while incremental learning increases these values to 78.5%, 73.4% and 82.7%, respectively, after ten steps (Figure 2).

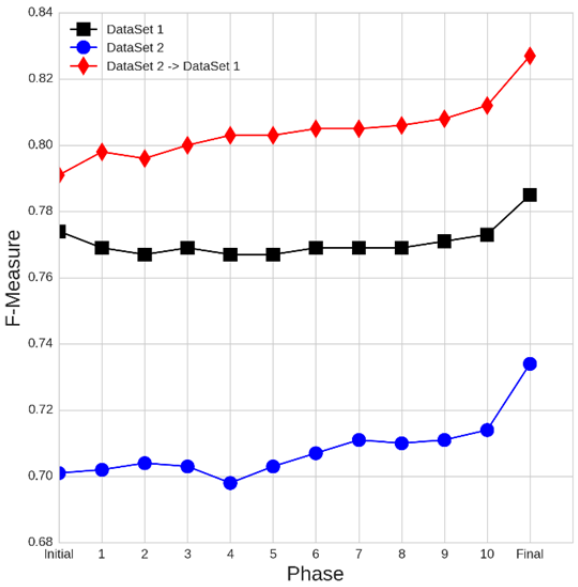

Figure 2. Accuracy gain from incremental learning with additional labeled data.

5. CONCLUSIONS

We provide a GUI-based technology that supports the prediction of standard sentiment classes and allows for a) relabeling predictions or adding labeled instances to retrain the weights of a given model, and b) customizing lexical resource to account for false positives and false negatives. The tool supports interactive result exploration and model adjustment. This might be particularly helpful for users from the computational social sciences and humanities, where distant reading techniques, e.g. sentiment analysis, are often combined with close reading techniques, i.e. zooming into selected relevant data points.

Our results show that updating given models with information from new, labeled instances increases accuracy. This approach allows users to account for changes in natural language use, such as the emergence of new terms and concepts, subtle changes in norms and vernacular, and cultural shifts; thereby reducing the risk of model accuracy deterioration over time.

6. ACKNOWLEDGEMENTS

This work is supported by Anheuser-Busch InBev. Marie Arends from AB InBev provided invaluable advice on this work. We thank the following people from UIUC: Liang Tao and Chieh-Li Chin for their help with technology development, and Jingxian Zhang and Aditi Khullar for their contributions to the technology.

7. REFERENCES

[1] Baccianella, S., Esuli, A., and Sebastiani, F., 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 2200-2204.

[2] Bottou, L., 1991. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*.

[3] Bottou, L., 1998. Online learning and stochastic approximations. *On-line learning in neural networks*, 1-34.

[4] Bottou, L., 2010. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of International Conference on Computational Statistics (COMPSTAT)*, Paris, France, 177-186.

[5] Diesner, J. and Evans, C., 2015. Little Bad Concerns: Using Sentiment Analysis to Assess Structural Balance in Communication Networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Short paper* (Paris, France2015).

[6] Glorot, X., Bordes, A., and Bengio, Y., 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. *Proceedings of the 28th International Conference on Machine Learning*, 513-520.

[7] Go, A., Bhayani, R., and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*.

[8] Grimmer, J. and Stewart, B.M., 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*, 3, 267-297.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter 11*, 1, 10-18. DOI= http://dx.doi.org/10.1145/1656274.1656278.

[10] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., and Potts, C., 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HTL)*, 142-150.

[11] Mohammad, S.M., Kiritchenko, S., and Zhu, X., 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)* Association for Computational Linguistics, 321-327.

[12] Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T., 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation* Association for Computational Linguistics, 312-320.

[13] Owoputi, O., O'Connor, B., Dyer, C., and Gimpel, K., 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, Atlanta, Georgia, USA, 380-390.

[14] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N., 2012. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. *Cmu-Ml-12-107*.

[15] Pang, B. and Lee, L., 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

[16] Pang, B., Lee, L., and Vaithyanathan, S., 2002. Thumbs up? In *ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)* Association for Computational Linguistics, Morristown, NJ, USA, 79-86. DOI= http://dx.doi.org/10.3115/1118693.1118704.

[17] Shanahan, J.G., Qu, Y., and Wiebe, J.M., 2006. *Computing attitude and affect in text: theory and applications*. Springer.

[18] Socher, R., Perelygin, A., and Wu, J., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1631-1642.

[19] Wilson, T., Wiebe, J., and Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT)* Association for Computational Linguistics, 347-354.

[20] Zhao, J., Dong, L., Wu, J., and Xu, K., 2012. MoodLens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, New York, New York, USA, 2-5.