Sanremo's winner is...
Category-driven Selection Strategies for Active Learning

Anne-Lyse Minard, Manuela Speranza, Mohammed R. H. Qwaider, Bernardo Magnini
Fondazione Bruno Kessler, Trento, Italy
{minard,manspera,qwaider,magnini}@fbk.eu

Abstract

English. This paper compares Active Learning selection strategies for sentiment analysis of Twitter data. We focus mainly on category-driven strategies, which select training instances taking into consideration the confidence of the system as well as the category of the tweet (e.g. positive or negative). We show that this combination is particularly effective when the performance of the system is unbalanced over the different categories. This work was conducted in the framework of automatically ranking the songs of "Festival di Sanremo 2017" based on sentiment analysis of the tweets posted during the contest.

Italiano. *Questo lavoro confronta strategie di selezione di Active Learning per l'analisi del sentiment dei tweet focalizzandosi su strategie guidate dalla categoria. Selezioniamo istanze di addestramento combinando la categoria del tweet (per esempio positivo o negativo) con il grado di confidenza del sistema. Questa combinazione è particolarmente efficace quando la distribuzione delle categorie non è bilanciata. Questo lavoro aveva come scopo il ranking delle canzoni del "Festival di Sanremo 2017" sulla base dell'analisi del sentiment dei tweet postati durante la manifestazione.*

1 Introduction

Active Learning (AL) is a well known technique for the selection of training samples to be annotated by a human when developing a supervised machine learning system. AL allows for the collection of more useful training data, while at the same time reducing the annotation effort (Cohn et al., 1994). In the AL framework samples are usually selected according to several criteria, such as informativeness, representativeness, and diversity (Shen et al., 2004).

This paper investigates AL selection strategies that consider the categories the current classifier assigns to samples, combined with the confidence of the classifier on the same samples. We are interested in understanding whether these strategies are effective, particularly when category distribution and category performance are unbalanced. By comparing several options, we show that selecting low confidence samples of the category with the highest performance is a better strategy than selecting high confidence samples of the category with the lowest performance.

The context of our study is the development of a sentiment analysis system that classifies tweets in Italian. We used the system to automatically rank the songs of Sanremo 2017 based on the sentiment of the tweets posted during the contest.

The paper is structured as follows. In Section 2 we give an overview of the state-of-the-art in selection strategies for AL. Then we present our experimental setting (Section 3) before detailing the tested selection strategies (Section 4). Finally, we describe the results of our experiment in Section 5 and the application of the system to ranking Sanremo's songs in Section 6.

2 Related Work

AL (Cohn et al., 1994; Settles, 2010) provides a well known methodology for reducing the amount of human supervision (and the corresponding cost) for the production of training datasets necessary in many Natural Language Processing tasks. An incomplete list of references includes Shen et al. (2004) for Named Entity Recognition, Ringger et al. (2007) for PoS Tagging, and Schohn and Cohn (2000) for Text Classification.

AL methods are based on strategies for sam-

ple selection. Although there are two main types of selection methods, certainty-based and committee-based, here we concentrate only on certainty-based selection methods. The main certainty-based strategy used is the uncertainty sampling method (Lewis and Gale, 1994). Shen et al. (2004) propose a strategy which is based on the combination of several criteria: informativeness, representativeness, and diversity. The results presented by Settles and Craven (2008) show that information density is the best criterion for sequence labeling. Tong and Koller (2002) propose three selection strategies that are specific to SVM learners and are based on different measures taking into consideration the distances to the decision hyperplane and margins.

Many NLP tasks suffer from unbalanced data. Ertekin et al. (2007) show that selecting examples within the margin overcomes the problem of unbalanced data.

The previously cited selection strategies are often applied to binary classification and do not take into account the predicted class. In this work we are interested in multi-class classification tasks, and in the problem of unbalanced data and dominant classes in terms of performance.

Esuli and Sebastiani (2009) define three criteria that they combine to create different selection strategies in the context of multi-label text classification. The criteria are based on the confidence of the system for each label, a combination of the confidence of each class for one document, and a weight (based on the F1-measure) assigned to each class to distinguish those for which the system performs badly. They show that in most of the cases this last criteria does not improve the selection.

Our applicative context is a bit different as we are not working on a multi-label task. Instead of computing a weight according to the F1-measure, we experimented with a change of strategy where we focus on a single class.

3 Experimental Setting

The context of our study was the development of a supervised sentiment analysis system that classifies tweets into one of the following four classes: `positive`, `negative`, `neutral`, and `n/a` (i.e. not applicable).

The manual annotation of the data was mainly performed by 25 3rd and 4th year students from local high schools who were doing a one-week group internship at Fondazione Bruno Kessler.

We created an initial training set using an AL mechanism that selects the samples with the lowest system confidence[1], i.e. those closer to the hyperplane and therefore most difficult to classify. In the following we describe the sentiment analysis system, the Active Learning process and the creation of the test and the initial training set. Finally, we introduce the experiments performed on selection strategies for Active Learning.

Sentiment Analysis System. Our system for sentiment analysis is based on a supervised machine learning method using the SVM-MultiClass tool (Joachims et al., 2009)[2]. We extract the following features from each tweet: the tokens composing the tweet, and the number of urls, hashtags, and aliases it contains. It takes as input a tokenized tweet[3] and returns as output its polarity.

AL Process. We used TextPro-AL, a platform which integrates an NLP pipeline, an AL mechanism and an annotation interface (Magnini et al., 2016). The AL process is as follows: (i) a large unlabeled dataset is annotated by the sentiment analysis system (with a small temporary model used to initialize the AL process[4]); (ii) samples are selected according to a selection strategy; (iii) annotators annotate the selected tweets; (iv) the new annotated samples are accumulated in the batch; (v) when the batch is full the annotated data are added to the existing training dataset and a new model is built; (vi) the unlabeled dataset is annotated again using the newly built model and the cycle begins again at (ii).

The unlabeled dataset consists of 400,000 tweets that contained the hashtag #Sanremo2017. The maximum size of the batch is 120, so retraining takes place every 120 annotated tweets.

Training and Performance. The initial training set, whose creation required half a day of work[5], is

[1]The confidence score is computed as the average of the margin estimated by the SVM classifier for each entity.

[2]`https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html`

[3]Tokenization is performed using the Twokenizer java library `https://github.com/vinhkhuc/Twitter-Tokenizer/blob/master/src/Twokenizer.java`

[4]The temporary model has been built using 155 tweets annotated manually by one annotator. After the first step of the AL process, these tweets are removed from the training set.

[5]The 25 high schools students worked in pairs and trios, for a total of 12 groups.

composed of 2,702 tweets. The class negative is the most represented, covering almost 40% of the total, with respect to positive, with around 30% of the total. The distribution of the two minor classes is rather close, with 18% for neutral and 13% for n/a.

As a test set we used 1,136 tweets randomly selected from among all the tweets which mentioned either a Sanremo song or singer. The test set was annotated partly by the high school students (656 tweets) and partly by two expert annotators (480 tweets); each tweet was annotated with the same category by at least two annotators. 58% of the tweets are positive, 20% are negative, 14% are neutral, and 8% are n/a.

We built the test set selecting the tweets randomly from the unlabeled dataset in order to make it representative of the whole dataset.

The overall performance of the system trained on the initial set is 40.7 in terms of F1 (see EVAL2702 in Table 1). The F1 obtained on the two main categories, i.e. positive and negative, is 54.5, but the system performs more poorly on negative than on positive, with F1-measures of 33.6 and 75.4 respectively.

Experiment. As the evaluation showed good results on positive but poor results on negative, we devised and tested novel selection strategies better able to balance the performance of the system over the two classes. We divided the 25 annotators into three different groups: each group annotated 775 tweets. The tweets annotated by the first group were selected with the same strategy used before, whereas for the other two groups we implemented two new selection strategies taking into account not only the confidence of the system but also the class it assigns to a tweet. As a result we obtained three different extensions of the same size and were thus able to compare the performance of the system trained on the initial training set plus each of the extensions.

4 Selection Strategies

We tested three selection strategies that take into account the classification proposed by the system in order to select the most useful samples to improve the distinction between positive and negative.

S1: low confidence. The first strategy we tested is the baseline strategy, which selects tweets classified by the system with the lowest confidence. The low confidence strategy was also used to build the initial training set (S0: lowC) as described is Section 3.

S2: NEGATIVE with high confidence. The second strategy consists of selecting the samples classified as negative with the highest confidence. We assume that this will increase the amount of negative tweets selected, thus enabling us to improve the performance of the system on the negative class. Nevertheless, as the system has a high confidence on the classification of these tweets, through this strategy we are adding easy examples to the training set that the system is probably already able to classify correctly.

S3: POSITIVE with low confidence. The third strategy aims at selecting the positive tweets for which the system has the lowest confidence. We expect in this way to get the difficult cases, i.e. tweets that are close to the hyperplane and that are classified as positive but whose classification has a high chance of being incorrect.

As the initial system has high recall (82.8) but low precision (69.3) for the class positive, we assume that it needs to improve on the examples wrongly classified as positive. We expect that inside the tweets wrongly classified as positive we will find difficult cases of negative tweets which will help to improve the system on the negative class. On the other hand, recall for the negative class is low (25.7), whereas precision is slightly better (48.7), which is why we decided to extract positive tweets with low confidence instead of negative tweets with low confidence.

5 Results and Discussion

In Table 1 we present the results (in tersm of F1) obtained by the system using the additional training data selected through the three different selection strategies described above. In order to facilitate the interpretation of the results, we also report the performance obtained by the system trained only on the initial set of 2,702 tweets. Additionally, in Table 2, we give the results obtained by the system for each configuration also in terms of recall and precision (besides F1).

The first four lines report the results for each of the four categories, while lines six and seven report respectively the macro-average F1 over the four classes and the macro-average F1 over the

		Eval2702		Experiment on selection strategies					
Strategy used		S0: lowC		S1: lowC		S2: NEG-highC		S3: POS-lowC	
		F1	tweets	F1	tweets	F1	tweets	F1	tweets
NEGATIVE		33.6	1,080	34.8	1,374	32.0	1,669	39.3	1,299
	wrt S0	-	-	(+1.2)	(+294)	(-1.6)	(+589)	(+5.7)	(+219)
POSITIVE		75.4	798	74.8	975	74.8	869	76.5	1,065
	wrt S0	-	-	(-0.6)	(+177)	(-0.6)	(+71)	(+1.1)	(+267)
NEUTRAL		22.3	476	20.9	595	23.3	567	24.6	672
	wrt S0	-	-	(-1.4)	(+119)	(+1.0)	(+91)	(+2.3)	(+196)
N/A		31.3	348	28.6	533	27.6	372	28.6	441
	wrt S0	-	-	(-2.7)	(+185)	(-3.7)	(+24)	(-2.7)	(+93)
Average 4 classes		40.7	2,702	39.8	3,477	39.4	3,477	42.3	3,477
	wrt S0	-	-	(-0.9)	(+775)	(-1.3)	(+775)	(+1.6)	(+775)
Average POS/NEG		54.5	-	54.8	-	53.4	-	57.9	-
	wrt S0	-	-	(+0.3)	-	(-1.1)	-	(+3.4)	-

Table 1: Performance of the system trained on 2,702 tweets and performance of the system trained on the same set of data incremented with 775 tweets selected through three different selection strategies.

	Eval2702			Experiment on selection strategies								
Strategy used	S0: lowC			S1: lowC			S2: NEG-highC			S3: POS-lowC		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
NEGATIVE	25.7	**48.7**	33.6	28.4	45.0	34.8	24.3	46.6	32.0	30.6	54.8	39.3
POSITIVE	**82.8**	69.3	75.4	81.6	69.0	74.8	82.2	68.7	74.8	**85.3**	69.3	76.5
NEUTRAL	20.1	**25.0**	22.3	17.7	25.4	20.9	20.7	26.6	23.3	21.3	**29.2**	24.6
N/A	**32.6**	30.0	31.3	30.4	26.9	28.6	29.3	26.0	27.6	27.2	30.1	28.6
Average 4 classes	40.3	43.2	40.7	39.5	41.6	39.8	39.2	41.9	39.4	41.1	45.9	42.3
Average POS/NEG	54.3	59.0	54.5	55.0	57.0	54.8	53.3	57.6	53.4	57.9	62.1	57.9

Table 2: Performance in terms of precision, recall and F1 of the system trained on the different training set. The two last lines are the average of the recall, precision and F1 over 4 and 2 classes.

two most important classes, i.e. `positive` and `negative`. For each selection strategy, we indicate the difference in performance obtained with respect to the system trained on the initial set, as well as the number of annotated tweets that have been added.

With the baseline strategy (S1: lowC, i.e., selection of the tweets for which the system has the lowest confidence) the performance of the system decreases slightly, from an F1 of 40.7 to an F1 of 39.8. Most of the added samples are negative tweets (38%), which enables the system to increase its performance on this class by 1.2 points.

When using the second strategy (S2: NEG-highC, i.e. selection of the negative tweets with the highest confidence), 76% of the new tweets are negative, but the performance of the system on this class decreases. Even the overall performance of the system decreases, despite adding 775 tweets.

We observe that the best strategy is S3 (POS-lowC, i.e., selection of the positive tweets with the lowest confidence), with an improvement of the macro-average F1-measure over the 4 classes by 1.6 points and over the `positive` and `negative` classes by 3.4 points. Although we add more positive than negative tweets to the training data (34%), the performance of the system on the `negative` class increases as well, from F1 33.6 to F1 39.3. This strategy worked very well in enabling us to select the examples which help the system discriminate between the two main classes.

6 Application: Sanremo's Ranking

After evaluating the three different selection strategies, we trained a new model using all the tweets that had been annotated. With this new model, as expected, we obtained the best results. The average F-measure on the `negative` and

positive classes is 58.2, the average F-measure over the 4 classes is 42.1.

For the annotation to be used for producing the automatic ranking, we provided the system with some gazetteers, i.e. a list of words that carry positive polarity and a list of words that carry negative polarity. We thus obtained a small improvement in system performance, with an F1 of 42.8 on the average of the four classes and an F1 of 58.3 on the average of positive and negative.

As explained in the Introduction, the applicative scope of our work was to rank the songs competing in Sanremo 2017. For this, we used only the total number of tweets talking about each singer and the polarity assigned to each tweet by the system. In total we had 118,000 tweets containing either a reference to a competing singer or song that had been annotated automatically by the sentiment analysis system. By doing the ranking according to the proportion of positive tweets of each singer, we were able to identify 4 out of the top 5 songs and 4 out of the 5 last place songs. In Table 3, we show the official ranking versus the automatic ranking. The Spearman's rank correlation coefficient between the official ranking and our ranking is 0.83, and the Kendall's tau coefficient is 0.67

Singer	Official	System
Francesco Gabbani	1	8
Fiorella Mannoia	2	4
Ermal Meta	3	1
Michele Bravi	4	2
Paola Turci	5	5
Sergio Sylvestre	6	6
Fabrizio Moro	7	3
Elodie	8	9
Bianca Atzei	9	13
Samuel	10	7
Michele Zarrillo	11	10
Lodovica Comello	12	12
Marco Masini	13	14
Chiara	14	11
Alessio Bernabei	15	16
Clementino	16	15

Table 3: Sanremo's official ranking and the ranking produced by our system

7 Conclusion

We have presented a comparative study of three AL selection strategies. We have shown that a strategy that takes into account both the automatically assigned category and the system's confidence performs well in the case of unbalanced performance over the different classes.

To complete our study it would be interesting to perform further experiments on other multi-classification problems. Unfortunately this work required intensive annotation work and so its replication on other tasks would be very expensive. A lot of work on Active Learning has been done using existing annotated corpora, but we think that it is too far from a real annotation situation as the datasets used are generally limited in tems of size.

In order to test different selection strategies, we have evaluated the sentiment analysis system against a gold standard, but we have also performed an application-oriented evaluation by ranking the songs participating in Sanremo 2017.

As future work, we want to explore the possibility of automatically adapting the selection strategies while annotating. For example, if the performance of the classifier of one class is low, the strategy in use could be changed in order to select the samples needed to improve on that class.

Acknowledgments

This work has been partially funded by the EuclipRes project, under the program *Bando Innovazione 2016* of the Autonomous Province of Bolzano. We also thank the high school students who contributed to this study with their annotation work within the FBK Junior initiative.

References

David Cohn, Richard Ladner, and Alex Waibel. 1994. Improving generalization with active learning. In *Machine Learning*, pages 201–221.

Seyda Ertekin, Jian Huang, Léon Bottou, and C. Lee Giles. 2007. Learning on the border: active learning in imbalanced data classification. In Mário J. Silva, Alberto H. F. Laender, Ricardo A. Baeza-Yates, Deborah L. McGuinness, Bjørn Olstad, Øystein Haug Olsen, and André O. Falcão, editors, *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 127–136. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2009. Active learning strategies for multi-label text classification. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soulé-Dupuy, editors, *Advances in Information Retrieval, 31th European*

Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings, volume 5478 of *Lecture Notes in Computer Science*, pages 102–113. Springer.

Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural svms. *Mach. Learn.*, 77(1):27–59, October.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proc.International ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 3–12, New York, NY, USA. Springer-Verlag New York, Inc.

Bernardo Magnini, Anne-Lyse Minard, Mohammed R. H. Qwaider, and Manuela Speranza. 2016. TEXTPRO-AL: An Active Learning Platform for Flexible and Efficient Production of Training Data for NLP Tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*.

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Greg Schohn and David Cohn. 2000. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 839–846, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1070–1079. ACL.

Burr Settles. 2010. Active learning literature survey. Technical report.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March.