Dynamic Parameter Adaptation of SVM Based Active Learning Methodology

Jasmina Smailović¹ Miha Grčar¹ Nada Lavrač^{1,2} M

Martin Žnidaršič¹

¹Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia (name.surname@ijs.si) ²University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

Abstract

In this paper we present experimental assessment of a dynamic adaptation of an approach for sentiment classification of tweets. Specifically, this approach enables a dynamic adaptation of the parameters used for three-class classification with a binary SVM classifier. The approach is suited for incremental active learning scenarios in domains with frequent concept alterations and changes. Our target application is in domain of finance and the assessment is partially domain-specific, but the approach itself is not limited to a particular domain.

1 Introduction

The work presented in this paper is aimed at the analysis of sentiment in Twitter messages, which became a very common and well studied problem [KZM14, MGS16, NRR⁺16]. Our specific focus, though, is on employment of techniques of incremental active learning in a financial domain. The general aim of our work is to develop a methodology that would allow keeping a stock-company focused sentiment classifier up-to-date with minimal human effort. Namely, in case of using informal data sources, like Twitter, and in dynamic target domains such as finance, updating of sentiment classifiers is necessary, as new data features emerge and existing features can change or even reverse their impact on sentiment classification. To take this into account, sentiment lexicon based approaches must update the lexicons, while in machine learning approaches that work with n-grams, the learning processes have to be repeated or an incremental learning algorithm must be employed. In any case, new labeled data is needed, which usually represents the main practical obstacle. Namely, labeling new data in this domain requires human expert effort, thus its frequency is limited (e.g., we cannot get hundreds of new labels per second, even if the cost would not be a constraint) and usually its volume (cost) as well. Therefore, it is beneficial to use an appropriate active learning strategy in order to limit this effort as much as possible.

A particular analysis of active learning strategies that we present in this paper is concerned with the assessment of impacts of a technique for dynamic adaptation of the parameters of an SVM based active learning. Specifically, we elaborate upon the concept of the dynamic neutral zone [Sma14] and present an extended experimental assessment of this approach. The neutral zone is the area around the SVM classifier's hyperplane that distinguishes among the examples that are to be classified as positive and those that are to be classified as negative [SGLŽ13, SGLŽ14, SKG⁺15]. Definition of such an area allows for three-class classification (negative, neutral, positive) with a binary SVM classifier in cases when only positive and negative learning data is available. An adaptive version of such an area definition, which was recently proposed [Sma14] and denoted as dynamic neutral zone, is able to adapt to the characteristics of new labeled data that becomes available by active learning.

 $[\]label{eq:copyright} \textcircled{C} \ by \ the \ paper's \ authors. \ Copying \ permitted \ for \ private \ and \ academic \ purposes. \\$

In: G. Krempl, V. Lemaire, E. Lughofer, and D. Kottke (eds.): Proceedings of the Workshop Active Learning: Applications, Foundations and Emerging Trends, AL@iKNOW 2016, Graz, Austria, 18-OCT-2016, published at http://ceur-ws.org

The presented active learning methodology and the approach for examining the relationship between tweet sentiment and stock prices is explained in detail in our previous studies [Sma14, SGLŽ13, SGLŽ14]. In Section 2 we briefly revisit the active learning approach and the concept of the neutral zone. The new extended experiments with the dynamic neutral zone are listed in Section 3 and discussed in Section 4.

2 Methodology

The aim of our experiments is to discover the best combination of parameters of the developed active learning methodology [Sma14, SGLŽ14] for sentiment analysis. The initial sentiment model is trained using the smiley-labeled Twitter messages [GBH09]¹ by employing the Support Vector Machine (SVM) [Vap95] algorithm. We measure the model performance (in terms of the F-measure of the positive class) on a simulated stream of tweets by employing the holdout evaluation approach adjusted for dynamic environments [BK09, IGD11], that is, we evaluate the model on each new batch of data from the Twitter data stream. The simulated data stream consists of tweets which discuss Baidu² stocks in year 2011. Moreover, active learning is performed, i.e. a selection of tweets from each batch is chosen to be manually labeled and added to the model.

The sentiment model is trained using the positive and negative tweets. However, in the classification phase we adjust the output of the SVM algorithm to detect also the neutral tweets by employing the concept of the neutral zone, that is, examples which are positioned in the neutral zone are marked as neutral. There are various ways of implementing the concept of the neutral zone. For example, the fixed neutral zone is constrained by empirically predefined boundaries [SGLŽ13, SGLŽ14, Sma14] as sketched in Figure 1(a), while the relative neutral zone is a function of positive and negative average distances of training examples [SKG⁺15, Sma14]. The key idea of the latter approach is the following. Given that an example is projected on the positive side of the SVM hyperplane at distance d and the average distance of positive training examples is \bar{d}_+ , the first step is to calculate the classification reliability by applying the following formula [SKG⁺15, Sma14]:

$$R = \frac{d}{2 * \bar{d}_+} \tag{1}$$

If the calculated reliability is greater than 1, it is transformed to R = 1. The example is labeled as neutral if its classification reliability is below a predefined reliability threshold R_T . Figure 1(b) presents an example of classifying an instance (at distance d) in this setting. The same approach (with using the average distances of negative training examples) is applied if an example is projected on the negative side of the SVM hyperplane.

In the active learning environment we dynamically update not only the sentiment model, but also the parameters of the relative neutral zone, i.e. the average training distances. The positive average distance is updated by applying the following formula [Sma14]:

$$\bar{d_{+}}' = (1 - \alpha) * \bar{d}_{+} + \alpha * \bar{d}_{b}$$
 (2)

where $\bar{d_+}'$ is an updated average distance, $\bar{d_+}$ is the current one, and $\bar{d_b}$ is the average distance of the positive examples in the currently processed batch *b*, which were used for updating the model. Parameter α controls the influence of the new and previous tweets. If α is set to 0, the average distance of initial training examples does not get dynamically updated. Equation 2 is applied accordingly for dynamically updating the negative average distance.

We experimented with the following active learning query strategies [Sma14, SGLŽ14] to select the most suitable examples from each batch of data for manual labeling:

- 1. Closest to the neutral zone: the algorithm chooses a selection of tweets whose classification reliability is closest to the reliability threshold. The number of positive/negative examples (according to the classifier's labeling) must not exceed half of the allocated manual labels.
- 2. Random: the algorithm randomly selects tweets for manual labeling.
- 3. Combined approach: combination of two previous approaches, i.e. a certain percentage of tweets is chosen randomly, while the rest of the tweets are chosen according to the "Closest to the neutral zone" strategy.

¹The dataset was obtained from the Sentiment140 Web page, section "For Academics" (http://help.sentiment140.com/for-students).

²http://www.baidu.com/.

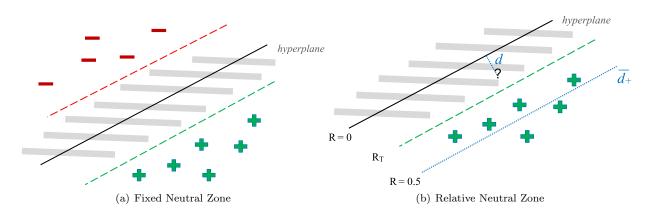


Figure 1: Two Approaches to the Concept of the Neutral Zone

Additionally, we evaluated the scenario without active learning, i.e. without updating the sentiment model or neutral zone.

The Friedman test [Dem06, Fri37, Fri40], the Iman-Davenport improvement [ID80], and the Nemenyi posthoc test [Nem63] were used to rank a selection of the evaluated active learning settings and to find statistically significant differences between them.

The implementation of the methodology uses elements of several libraries: Pegasos SVM [SSSS07] from the sofia-ml library³ [Scu10], SWIG⁴ to connect sofia-ml C++ implementation with C# programming language, and the LATINO library⁵ for preparing the features. The learning algorithm for the initial model training in sofia-ml was adapted by implementing sampling which takes examples in succession.

3 Experiments

In this study we extend the experimental setting from [Sma14] and test the following parameters and their values (the parameters used already in [Sma14] are also included):

- Alpha values: 0, 0.05, 0.1, 0.3, 0.5.
- Five active learning querying strategies and one without active learning.
- Two batch selection strategies: select 10 of 100 (select 10 examples for manual labeling out of 100 examples in a batch) and select 10 of 50.
- Reliability threshold: 0, 0.1, 0.2, 0.3, 0.4, 0.5.

The results of experimental assessment for all combinations of the above parameters are presented in Table 1. From Table 1 it is not straightforward to conclude which combination of parameters is the best one. For that reason, we present the average results of active learning strategies over all values of reliability threshold in Figure 2, where lighter color corresponds to lower values and darker color corresponds to higher (better) values. We exclude the "AL closest to NZ" strategy since the results in Table 1, which are marked with asterisk(s), indicate that this strategy is unreliable as many batches did not have positively classified tweets, which caused missing values of F-measure (see [Sma14] for more details on this phenomenon). The Figure 2 indicates that both 0.1 and 0.3 are reasonable values for the parameter α . However, we focus on $\alpha = 0.3$, since we already performed the analysis of $\alpha = 0.1$ in our previous study [Sma14].

³https://code.google.com/p/sofia-ml/.

⁴http://www.swig.org/.

⁵https://github.com/LatinoLib/LATINO.

Table 1: Average F-measure of the Positive Class \pm Std. Deviation
for Different Active Learning and Batch Selection Strategies, $Alpha$
Values, and Reliability Thresholds

			$\alpha = 0$				
Rel. threshold	0	0.1	0.2	0.3	0.4	0.5	
Select 10 of 100							
AL closest to NZ	$0.5512 {\pm} 0.12$	$0.5396 {\pm} 0.12$	0.5282 ± 0.11	$0.5165 {\pm} 0.11$	$0.5018 {\pm} 0.11$	0.4802 ± 0.11	
AL comb. 20% r.	$0.5512{\pm}0.12$	$0.5396{\pm}0.12$	$0.5281 {\pm} 0.11$	$0.5164{\pm}0.11$	$0.5017 {\pm} 0.11$	$0.4803{\pm}0.11$	
AL comb. 50% r.	$0.5513 {\pm} 0.12$	$0.5398 {\pm} 0.12$	$0.5283 {\pm} 0.11$	$0.5165 {\pm} 0.11$	5165 ± 0.11 0.5016 ± 0.11		
AL comb. 80% r.	$0.5512{\pm}0.12$	$0.5396{\pm}0.12$	$0.5281 {\pm} 0.11$	$0.5165 {\pm} 0.11$	$0.5017 {\pm} 0.11$	$0.4803 {\pm} 0.11$	
AL 100% rand.	$0.5514{\pm}0.12$	$0.5399 {\pm} 0.12$	$0.5281 {\pm} 0.11$	$0.5169 {\pm} 0.11$	$0.5017 {\pm} 0.11$	$0.4804{\pm}0.11$	
No AL	$0.5500 {\pm} 0.12$	$0.5389 {\pm} 0.12$	$0.5277 {\pm} 0.11$	$0.5162{\pm}0.11$	$0.5004{\pm}0.11$	$0.4787 {\pm} 0.10$	
Select 10 of 50							
AL closest to NZ	$0.5466 {\pm} 0.14$	$0.5342 {\pm} 0.14$	0.5221 ± 0.14	0.5103 ± 0.14	$0.4956 {\pm} 0.13$	$0.4756 {\pm} 0.13$	
AL comb. 20% r.	$0.5466 {\pm} 0.14$	$0.5339 {\pm} 0.14$	$0.5220 {\pm} 0.14$	$0.5103 {\pm} 0.14$	$0.4957 {\pm} 0.13$	$0.4757 {\pm} 0.13$	
AL comb. 50% r.	$0.5465 {\pm} 0.14$	$0.5340{\pm}0.14$	$0.5219 {\pm} 0.14$	$0.5104{\pm}0.14$	$0.4957 {\pm} 0.13$	$0.4758 {\pm} 0.13$	
AL comb. 80% r.	$0.5468 {\pm} 0.14$	$0.5340{\pm}0.14$	$0.5218 {\pm} 0.14$	$0.5103 {\pm} 0.14$	$0.4959 {\pm} 0.13$	$0.4756 {\pm} 0.13$	
AL 100% rand.	$0.5466 {\pm} 0.14$	$0.5341{\pm}0.14$	$0.5222 {\pm} 0.14$	$0.5109 {\pm} 0.14$	$0.4963{\pm}0.13$	$0.4762 {\pm} 0.13$	
No AL	$0.5444 {\pm} 0.14$	$0.5329 {\pm} 0.14$	0.5213 ± 0.14	$0.5094{\pm}0.14$	$0.4938 {\pm} 0.13$	$0.4731 {\pm} 0.13$	
			$\alpha = 0.05$				
Rel. threshold	0	0.1	0.2	0.3	0.4	0.5	
Select 10 of 100	Ť	0.2	0.2	0.0	0.1		
AL closest to NZ	$0.5512 {\pm} 0.12$	$0.5484{\pm}0.12$	$0.5749 \pm 0.11*$	* 0.5688±0.11*	* 0.5667+0.11*	* 0.5547+0.11**	
AL comb. 20% r.	0.5513 ± 0.12	0.5460 ± 0.12	0.5408 ± 0.12	0.5329 ± 0.11	0.5298 ± 0.11	0.5024 ± 0.11	
AL comb. 50% r.	0.5512 ± 0.12	0.5424 ± 0.12	0.5344 ± 0.11	0.5221 ± 0.11	0.5084 ± 0.11	0.4545 ± 0.10	
AL comb. 80% r.	0.5515 ± 0.12	0.5377 ± 0.11	0.5250 ± 0.11	0.5126 ± 0.11	0.4821 ± 0.11	0.4435 ± 0.11	
AL 100% rand.	0.5510 ± 0.12 0.5514 ± 0.12	0.5340 ± 0.11	0.5168 ± 0.11	0.4971 ± 0.11	0.4640 ± 0.11	0.4296 ± 0.11	
No AL	0.5500 ± 0.12	0.5389 ± 0.12	0.5277 ± 0.11	0.5162 ± 0.11	0.5004 ± 0.11	0.4787 ± 0.10	
Select 10 of 50	0.0000±0.12	0.0000±0.12	0.0211±0.11	0.0102±0.11	0.0001±0.11	0.1101±0.10	
AL closest to NZ	0.5471 ± 0.14	0.5613 ± 0.14	$0.5394{\pm}0.14$	0.5341 ± 0.14	$0.5985\pm0.12*$	* 0.5920±0.12**	
AL comb. 20% r.	0.5466 ± 0.14	0.5406 ± 0.14	0.5365 ± 0.14	0.5278 ± 0.14	0.5243 ± 0.14	0.5126 ± 0.14	
AL comb. 50% r.	0.5465 ± 0.14	0.5367 ± 0.14	0.5258 ± 0.14	0.5155 ± 0.14	0.4939 ± 0.13	0.4518 ± 0.13	
AL comb. 80% r.	0.5467 ± 0.14	0.5331 ± 0.14	0.5280 ± 0.11 0.5185 ± 0.14	0.4979 ± 0.13	0.4713 ± 0.13	0.4445 ± 0.13	
AL 100% rand.	0.5466 ± 0.14	0.5295 ± 0.14	0.5149 ± 0.14	0.4961 ± 0.13	0.4740 ± 0.13	0.4491 ± 0.13	
No AL	0.5444 ± 0.14	0.5329 ± 0.14	0.5213 ± 0.14	0.5094 ± 0.14	0.4938 ± 0.13	0.4731 ± 0.13	
	0.0111±0.11	0.0020±0.11	$\frac{0.0210\pm0.11}{\alpha=0.1}$	0.0001±0.11	0.1000±0.10	0.1101±0.10	
Rel. threshold	0	0.1	$\frac{\alpha - 0.1}{0.2}$	0.3	0.4	0.5	
Select 10 of 100	0	0.1	0.2	0.0	1.0	0.0	
AL closest to NZ	0.5512 ± 0.12	$0.5808 \pm 0.12*$	* 0 5800+0 10*	* 0.5923±0.10*	* 0 5356+0 11*	* 0 5765+0 10**	
AL comb. 20% r.				0.5325 ± 0.10 0.5375 ± 0.11			
AL comb. 20% r.	0.5530 ± 0.12 0.5513 ± 0.12	0.5405 ± 0.12 0.5415 ± 0.12	0.5320 ± 0.11 0.5320 ± 0.12	0.5246 ± 0.11	0.5289 ± 0.11 0.5116 ± 0.11	0.3102 ± 0.11 0.4831 ± 0.11	
AL comb. 50% r.	0.5513 ± 0.12 0.5512 ± 0.12	0.5415 ± 0.12 0.5384 ± 0.11	0.5279 ± 0.12	0.5240 ± 0.11 0.5128 ± 0.11	0.4833 ± 0.11	0.4031 ± 0.11 0.4531 ± 0.10	
AL 100% rand.	0.5512 ± 0.12 0.5514 ± 0.12	0.5335 ± 0.11 0.5335 ± 0.11	0.5279 ± 0.11 0.5164 ± 0.11	0.3123 ± 0.11 0.4961 ± 0.11	0.4638 ± 0.11 0.4638 ± 0.11	0.4323 ± 0.10 0.4323 ± 0.11	
No AL	0.5514 ± 0.12 0.5500 ± 0.12	0.5339 ± 0.11 0.5389 ± 0.12	0.5104 ± 0.11 0.5277 ± 0.11	0.4901 ± 0.11 0.5162 ± 0.11	0.4038 ± 0.11 0.5004 ± 0.11	0.4323 ± 0.11 0.4787 ± 0.10	
Select 10 of 50	0.00010.12	0.0009±0.12	0.0211±0.11	0.0102±0.11	0.0004±0.11	0.4707±0.10	
AL closest to NZ	$0.5766\pm0.15*$	* 0.5682±0.14*	$0.6340\pm0.12*$	* 0.6348±0.11*	* 0 6250+0 11**	* 0 5114+0 14*	
AL comb. 20% r.	0.5760 ± 0.13 0.5466 ± 0.14	0.5082 ± 0.14 0.5398 ± 0.14	0.0349 ± 0.12 0.5382 ± 0.14	0.0348 ± 0.11 0.5328 ± 0.14	0.5230 ± 0.11 0.5237 ± 0.14	0.5114 ± 0.14 0.5172 ± 0.14	
AL comb. 20% r.	0.5460 ± 0.14 0.5464 ± 0.14	0.5398 ± 0.14 0.5359 ± 0.14	0.5382 ± 0.14 0.5262 ± 0.14		0.3237 ± 0.14 0.4957 ± 0.14	0.3172 ± 0.14 0.4690 ± 0.13	
AL comb. 30% r. AL comb. 80% r.	0.5464 ± 0.14 0.5463 ± 0.14	0.5309 ± 0.14 0.5303 ± 0.14	0.5262 ± 0.14 0.5188 ± 0.14	0.5173 ± 0.14 0.5020 ± 0.14	0.4957 ± 0.14 0.4756 ± 0.13	0.4090 ± 0.13 0.4359 ± 0.13	
AL cond. 80% r. AL 100% rand.				$0.5020 {\pm} 0.14$ $0.4967 {\pm} 0.14$	0.4750 ± 0.13 0.4751 ± 0.13	0.4539 ± 0.13 0.4521 ± 0.13	
No AL	$0.5466 {\pm} 0.14$ $0.5444 {\pm} 0.14$	0.5299 ± 0.14 0.5329 ± 0.14	$0.5153 {\pm} 0.14$ $0.5213 {\pm} 0.14$	0.4907 ± 0.14 0.5094 ± 0.14	0.4751 ± 0.13 0.4938 ± 0.13	0.4521 ± 0.13 0.4731 ± 0.13	
INO AL	0.0444±0.14	0.0029±0.14	0.0210±0.14	0.0094±0.14	0.4990±0.19	0.4101±0.10	

			$\alpha = 0.3$					
Rel. threshold	0	0.1	0.2	0.3	0.4	0.5		
Select 10 of 100								
AL closest to NZ	$0.5499 \pm 0.12^*$	$0.5734 \pm 0.11^{**}$	* 0.5394±0.11**	* 0.5320±0.12**	0.5322 ± 0.11	$0.6219 \pm 0.09^{**}$		
AL comb. 20% r.	$0.5617 {\pm} 0.12$	$0.5442{\pm}0.11$	$0.5533 {\pm} 0.11$	$0.5420{\pm}0.11$	$0.5356{\pm}0.11$	$0.5244{\pm}0.11$		
AL comb. 50% r.	$0.5497{\pm}0.12$	$0.5431{\pm}0.12$	$0.5284{\pm}0.11$	$0.5160{\pm}0.12$	$0.5173 {\pm} 0.11$	$0.4806{\pm}0.12$		
AL comb. 80% r.	$0.5554{\pm}0.12$	$0.5356{\pm}0.11$	$0.5311{\pm}0.11$	$0.5022{\pm}0.11$	$0.4817 {\pm} 0.11$	$0.4639{\pm}0.12$		
AL 100% rand.	$0.5514{\pm}0.12$	$0.5337 {\pm} 0.11$	$0.5161 {\pm} 0.11$	$0.4951{\pm}0.11$	$0.4701 {\pm} 0.11$	$0.4337 {\pm} 0.12$		
No AL	$0.5500{\pm}0.12$	$0.5389{\pm}0.12$	$0.5277 {\pm} 0.11$	$0.5162{\pm}0.11$	$0.5004{\pm}0.11$	$0.4787 {\pm} 0.10$		
Select 10 of 50								
AL closest to NZ	$0.5557 {\pm} 0.15^*$	$0.5307 \pm 0.15^*$	$0.5293 \pm 0.14^*$	$0.5210{\pm}0.14^*$	$0.5298 {\pm} 0.15^*$	$0.5124{\pm}0.14$		
AL comb. 20% r.	$0.5470 {\pm} 0.14$	$0.5398{\pm}0.14$	$0.5446{\pm}0.14$	$0.5305 {\pm} 0.14$	$0.5193{\pm}0.14$	$0.5162{\pm}0.13$		
AL comb. 50% r.	$0.5467 {\pm} 0.14$	$0.5348{\pm}0.14$	$0.5267 {\pm} 0.14$	$0.5230{\pm}0.14$	$0.5029 {\pm} 0.14$	$0.4654{\pm}0.14$		
AL comb. 80% r.	$0.5467 {\pm} 0.14$	$0.5343{\pm}0.14$	$0.5199{\pm}0.14$	$0.5035 {\pm} 0.14$	$0.4703 {\pm} 0.14$	$0.4411 {\pm} 0.14$		
AL 100% rand.	$0.5455 {\pm} 0.14$	$0.5291{\pm}0.14$	$0.5134{\pm}0.14$	$0.4922{\pm}0.14$	$0.4730{\pm}0.13$	$0.4517 {\pm} 0.14$		
No AL	$0.5444{\pm}0.14$	$0.5329 {\pm} 0.14$	$0.5213{\pm}0.14$	$0.5094{\pm}0.14$	$0.4938 {\pm} 0.13$	$0.4731{\pm}0.13$		
			$\alpha = 0.5$					
Rel. threshold	0	0.1	0.2	0.3	0.4	0.5		
Select 10 of 100								
AL closest to NZ	$0.5706 {\pm} 0.11$	$0.5442 \pm 0.12^*$	$0.5308 \pm 0.12^{**}$	$0.5251 \pm 0.11^{*}$	$0.5058 \pm 0.11^{**}$	$0.5299 \pm 0.12^{**}$		
AL comb. 20% r.	$0.5577 {\pm} 0.11$	$0.5506{\pm}0.12$	$0.5466 {\pm} 0.11$	$0.5427{\pm}0.12$	$0.5282{\pm}0.12$	$0.4902{\pm}0.12$		
AL comb. 50% r.	$0.5496{\pm}0.12$	$0.5381{\pm}0.11$	$0.5310{\pm}0.11$	$0.5253{\pm}0.11$	$0.5017 {\pm} 0.12$	$0.4488 {\pm} 0.13$		
AL comb. 80% r.	$0.5479 {\pm} 0.12$	$0.5392{\pm}0.12$	$0.5171 {\pm} 0.12$	$0.4998 {\pm} 0.11$	$0.4629 {\pm} 0.12$	$0.4539{\pm}0.11$		
AL 100% rand.	$0.5514{\pm}0.12$	$0.5339{\pm}0.11$	$0.5178 {\pm} 0.11$	$0.4941{\pm}0.11$	$0.4693 {\pm} 0.12$	$0.4362{\pm}0.12$		
No AL	$0.5500{\pm}0.12$	$0.5389{\pm}0.12$	$0.5277 {\pm} 0.11$	$0.5162{\pm}0.11$	$0.5004{\pm}0.11$	$0.4787 {\pm} 0.10$		
Select 10 of 50								
AL closest to NZ	$0.5414 \pm 0.14^*$	$0.5234{\pm}0.14^*$	$0.5101 \pm 0.14^*$	$0.5043 \pm 0.15^{*}$	$0.5299 {\pm} 0.14$	$0.4995 {\pm} 0.14$		
AL comb. 20% r.	$0.5370{\pm}0.15$	$0.5359{\pm}0.14$	$0.5336{\pm}0.14$	$0.5269 {\pm} 0.14$	$0.5188 {\pm} 0.14$	$0.5076 {\pm} 0.14$		
AL comb. 50% r.	$0.5463{\pm}0.14$	$0.5356{\pm}0.14$	$0.5287 {\pm} 0.14$	$0.5109 {\pm} 0.14$	$0.4998 {\pm} 0.14$	$0.4673 {\pm} 0.14$		
AL comb. 80% r.	$0.5486{\pm}0.14$	$0.5304{\pm}0.14$	$0.5185 {\pm} 0.14$	$0.5008 {\pm} 0.13$	$0.4692{\pm}0.14$	$0.4397 {\pm} 0.15$		
AL 100% rand.	$0.5440{\pm}0.14$	$0.5273 {\pm} 0.14$	$0.5108 {\pm} 0.14$	$0.4900 {\pm} 0.13$	$0.4696{\pm}0.13$	$0.4467 {\pm} 0.14$		
No AL	$0.5444{\pm}0.14$	$0.5329 {\pm} 0.14$	$0.5213 {\pm} 0.14$	$0.5094{\pm}0.14$	$0.4938 {\pm} 0.13$	$0.4731{\pm}0.13$		

** sample contains less than 50% of data batches from which positive F-measure could be calculated.

* sample contains less than 70% of data batches from which positive F-measure could be calculated.

The results of the Friedman test with the Iman-Davenport improvement and the Nemenyi post-hoc test for $\alpha = 0.3$ are presented in Figure 3. The strategies which are not significantly different are connected with a red line. From the figure it follows that the best active learning settings are: "Select 10 of 100 with AL comb. 20% random", "Select 10 of 50 with AL comb. 20% random" and "Select 10 of 100 with AL comb. 50% random".

Finally, we analyze the relationship between sentiment in tweets and stock closing prices of the discussed company. We apply the Granger causality test [Gra69] for different time lags and time periods on two time series: daily change of the positive sentiment probability and daily return in stock closing price [Sma14, SGLŽ13, SGLŽ14]. This statistical test indicates whether one time series is useful for predicting the values of another one. The results for top three active learning settings for $\alpha = 0.3$ are shown in Table 2. The significant results, after applying the Bonferroni correction [Abd07], are marked in bold (which corresponds to values lower than 0.025).

4 Discussion and conclusions

We presented an extended experimental assessment of the active learning methodology with dynamic neutral zone in which we were particularly interested in experimenting with the parameter α which dynamically updates the neutral zone as new examples arrive from the data stream. The conclusions of this extended study are in agreement with our previous one [Sma14]. The indications about the characteristics of the neutral zone are now strengthened, but for many aspects still lack a decisive statistical significance.

For $\alpha = 0$ all the active learning strategies are better than the strategy without active learning. However, the differences between the strategies are not so prominent (see Figure 2). On the other hand, the results with the dynamic neutral zone ($\alpha > 0$) are more diverse between different strategies and the new results in this setting show that besides completely random query strategy, also combinations with a strong random component (80%) are even worse than not applying active learning at all. In general, the "Select 10 of 100" batch selection seems to be somewhat better than "Select 10 of 50" selection, which is not intuitive, but might be partly caused by partitioning of the batches. Moreover, in larger batches the querying strategies might be more effective as they operate on larger number of different examples. Discovery of the exact cause would be a possible direction for further work. Regarding the query strategies, the combined seem to be the best ones, but the differences among them are usually not significant (see Figure 3). The Granger causality analysis showed that there is a relationship between sentiment in tweets and stock prices in specific time periods, mostly June-August, as already shown in [Sma14]. The relationship also depends on choosing an appropriate active learning setting and the value of reliability threshold.

Alpha	0	0.05	0.1	0.3	0.5
Select 10 of 100					
AL comb. 20% rand.	0.5196	0.5339	0.5365	0.5435	0.5360
AL comb. 50% rand.	0.5196	0.5188	0.5240	0.5225	0.5158
AL comb. 80% rand.	0.5196	0.5087	0.5111	0.5117	0.5035
AL 100% rand.	0.5197	0.4988	0.4989	0.5000	0.5005
No AL	0.5187	0.5187	0.5187	0.5187	0.5187
Select 10 of 50					
AL comb. 20% rand.	0.5140	0.5314	0.5331	0.5329	0.5267
AL comb. 50% rand.	0.5140	0.5117	0.5151	0.5166	0.5148
AL comb. 80% rand.	0.5141	0.5020	0.5015	0.5026	0.5012
AL 100% rand.	0.5144	0.5017	0.5026	0.5008	0.4981
No AL	0.5125	0.5125	0.5125	0.5125	0.5125

Figure 2: Averaged F-measure Results from Table 1 for All Active Learning Strategies (Except "AL Closest to NZ") Over All Values of Reliability Threshold

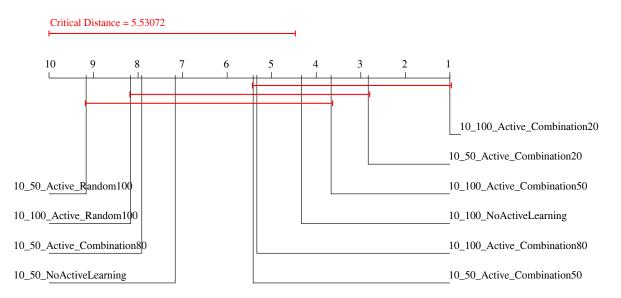


Figure 3: The Results of the Friedman Test With the Iman-Davenport Improvement and the Nemenyi Post-hoc Test for $\alpha = 0.3$

Table 2: Granger Causality Results (*p*-values) Between Daily Change of the Positive Sentiment Probability and Daily Return in Stock Closing Price for Baidu. The Results of Three Active Learning Query Strategies for $\alpha = 0.3$ are Shown. Statistically Significant Results are Marked in Bold

Reliability threshold	Lag	0	0.1	0.2	0.3	0.4	0.5
Select 10 of 100, comb. 20% rand.							
9 months	1	0.874	0.785	0.190	0.170	0.630	0.773
March - May	1	0.247	0.469	0.507	0.657	0.815	0.887
June - August	1	0.152	0.396	0.282	0.044	0.451	0.331
September - November	1	0.416	0.696	0.212	0.604	0.906	0.625
9 months	2	0.705	0.837	0.199	0.240	0.497	0.352
March - May	2	0.148	0.244	0.644	0.369	0.860	0.534
June - August	2	0.292	0.713	0.093	0.033	0.147	0.175
September - November	2	0.698	0.683	0.350	0.384	0.849	0.272
9 months	3	0.696	0.946	0.328	0.358	0.586	0.331
March - May	3	0.269	0.444	0.805	0.540	0.916	0.541
June - August	3	0.384	0.413	0.024	0.026	0.099	0.069
September - November	3	0.822	0.324	0.439	0.342	0.864	0.322
Select 10 of 50, comb. 20% rand.							
9 months	1	0.565	0.099	0.415	0.666	0.856	0.912
March - May	1	0.545	0.915	0.934	0.227	0.352	0.510
June - August	1	0.537	0.034	0.719	0.276	0.660	0.125
September - November	1	0.269	0.450	0.159	0.532	0.169	0.400
9 months	2	0.904	0.150	0.681	0.417	0.992	0.975
March - May	2	0.067	0.941	0.927	0.339	0.537	0.734
June - August	2	0.807	0.042	0.584	0.037	0.442	0.096
September - November	2	0.157	0.639	0.300	0.664	0.352	0.292
9 months	3	0.321	0.131	0.719	0.395	0.996	0.985
March - May	3	0.068	0.295	0.970	0.515	0.783	0.855
June - August	3	0.345	0.066	0.608	0.121	0.373	0.094
September - November	3	0.320	0.538	0.314	0.524	0.553	0.517
Select 10 of 100, comb. 50% rand.							
9 months	1	0.531	0.076	0.232	0.124	0.118	0.793
March - May	1	0.286	0.960	0.438	0.881	0.909	0.836
June - August	1	0.089	0.042	0.026	0.028	0.021	0.046
September - November	1	0.783	0.709	0.854	0.745	0.790	0.258
9 months	2	0.405	0.129	0.334	0.049	0.142	0.130
March - May	2	0.541	0.572	0.708	0.309	0.293	0.139
June - August	2	0.107	0.039	0.041	0.019	0.025	0.056
September - November	2	0.646	0.742	0.894	0.813	0.945	0.058
9 months	3	0.424	0.172	0.183	0.053	0.184	0.232
March - May	3	0.710	0.766	0.274	0.545	0.506	0.282
June - August	3	0.134	0.037	0.060	0.049	0.007	0.064
September - November	3	0.278	0.456	0.550	0.366	0.285	0.02

Acknowledgements

This work was partially funded by the European Commission in the context of the FP7 projects FIRST and FOC (Grant No. 257928 and 255987), by the Slovenian Research Agency through the research program Knowledge Technologies under (Grant P2-0103) and the project Influence of formal and informal corporate communications on capital markets (Grant No. J5-7387). We are grateful to Dragi Kocev for his help in the statistical evaluation of the results and Martin Saveski for his help with the implementation of the active learning algorithms.

References

- [Abd07] Herv Abdi. Bonferroni and Šidák corrections for multiple comparisons. In Neil Salkind, editor, Encyclopedia of Measurement and Statistics, pages 103–107. Thousand Oaks (CA): Sage, 2007.
- [BK09] Albert Bifet and Richard Kirkby. Data stream mining: A practical approach. 2009.
- [Dem06] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 7:1–30, 2006.
- [Fri37] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200):675–701, 1937.
- [Fri40] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 11(1):86–92, 1940.
- [GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1–12, 2009.
- [Gra69] Clive W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, pages 424–438, 1969.
- [ID80] Ronald L. Iman and James M. Davenport. Approximations of the critical region of the friedman statistic. Communications in Statistics-Theory and Methods, 9(6):571–595, 1980.
- [IGD11] Elena Ikonomovska, João Gama, and Sašo Džeroski. Learning model trees from evolving data streams. Data Mining and Knowledge Discovery, 23(1):128–168, 2011.
- [KZM14] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.
- [MGS16] Igor Mozetič, Miha Grčar, and Jasmina Smailović. Multilingual Twitter Sentiment Classification: The Role of Human Annotators. *PloS one*, 11(5):e0155036, 2016.
- [Nem63] Peter B. Nemenyi. Distribution-free Multiple Comparisons. PhD thesis, Princeton University, 1963.
- [NRR⁺16] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming), 2016.
- [Scu10] David Sculley. Combined regression and ranking. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 979–988. ACM, 2010.
- [SGLŽ13] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery* in Complex, Unstructured, Big Data, Lecture Notes in Computer Science Volume 7947, pages 77–88. Springer Berlin Heidelberg, 2013.
- [SGLŽ14] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203, 2014.
- [SKG⁺15] Jasmina Smailović, Janez Kranjc, Miha Grčar, Martin Žnidaršič, and Igor Mozetič. Monitoring the Twitter sentiment during the Bulgarian elections. In Proc. IEEE Intl. Conf. on Data Science and Advanced Analytics, pages 1–10. IEEE, 2015.
- [Sma14] Jasmina Smailović. Sentiment analysis in streams of microblogging posts. PhD thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, 2014.
- [SSSS07] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In Proceedings of the 24th International Conference on Machine Learning, pages 807–814, 2007.
- [Vap95] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.