Accepted for ISCBI 2018, http://www.iscbi.com/

Multilingual Sentiment Analysis for a Swiss Gig

Ela Pustulka-Hunt, Thomas Hanne, Eliane Blumer, Manuel Frieder
School of Business, University of Applied Sciences and Arts Northwestern Switzerland
Olten, Switzerland
elzbieta.pustulka@fhnw.ch, thomas.hanne@fhnw.ch, eliane.blumer@gmail.com, manuel.frieder@gmail.com

Abstract— **We are developing a multilingual sentiment analysis solution for a Swiss human resource company working in the gig sector. To examine the feasibility of using machine learning in this context, we carried out three sentiment assignment experiments. As test data we use 963 hand annotated comments made by workers and their employers. Our baseline, machine learning (ML) on Twitter, had an accuracy of 0.77 with the Matthews correlation coefficient (MCC) of 0.32. A hybrid solution, Semantria from Lexalytics, had an accuracy of 0.8 with MCC of 0.42, while a tenfold cross-validation on the gig data yielded the accuracy of 0.87, F1 score 0.91, and MCC 0.65. Our solution did not require language assignment or stemming and used standard ML software. This shows that with more training data and some feature engineering, an industrial strength solution to this problem should be possible.**

Keywords—sentiment analysis, machine learning application, natural language processing, gig economy

I. Introduction

Gig economy [1][15] poses new questions in the area of service quality assessment. One of those is how to provide relevant feedback to all transaction participants (the platform, the gig worker and the employer). Prediction of future participant behaviour and influencing it are also economically important. The first step in improving the relationship is understanding the current status, based on web feedback given by the participants, which is the reason for applying sentiment analysis techniques in this context [11][19][25].

In social media, sentiment analysis was recently shown to perform best when lexicon based methods are combined with machine learning [10] and found out that negative sentiment is harder to classify than positive sentiment. Earlier research suggested, however, that machine learning is better than lexical approaches [24]. The choice of methodology for a sentiment analysis task is an open question. We work in a real world scenario which requires very accurate sentiment assignment, so that at the end of a business day, we can find reliably all the negative comments and improve our business operations by addressing the concerns expressed by the partners using the platform. This consideration serves the needs of the platform provider directly, by reducing the number of problems that can arise in the future. It is also beneficial to the other parties who will experience better business conditions.

We have three parties: the company providing a business platform (gig work brokerage), a pool of companies or individuals who need labor, and the workforce (gig workers), see [26] for more context. After each gig, the employers and gig workers rate one another. If the business runs smoothly, we expect most comments to be positive or neutral. Negative comments, on the other hand, are to be identified reliably. The worker is required to enter a star rating (between⋆ and ⋆⋆⋆⋆) and the employer does the same. If the rating is low (1 to 2 stars), a comment is mandatory. There are, however, many comments with rating 3 to 4 stars which convey negative sentiment and the discrepancy between the sentiment and rating is to be further investigated. The gig platform currently communicates the star rating to the participating parties but does not feedback the comment itself. The analysis we perform will be ultimately fed back to the workers and employers in a digested form, which motivates the development of reliable automated sentiment assignment.

Our argument is structured as follows. In Section II we review related work, in Section III we present our methodology and, in Section IV, the results. Section V discusses our findings, Section VI outlines future work and Section VII concludes.

II. RELATED WORK

We review sentiment analysis (SA) for Twitter [30] and the performance of machine learning methods used in this context, as Twitter data is quite similar to ours. A very thorough review of Twitter SA can be found in a survey by Giachanou and Crestani [12]. Recent books in this area include [19] and [6]. We focus on the result quality, as our goal is to deliver and industrial-strength solution for gig work evaluation.

Sentiment analysis aims to classify the sentiment into one of the appropriate classes, for instance neutral, positive, negative, and mixed. The task to be automated is a classification task [20][28]. The classification methods used in this context include machine learning (ML), lexicon-based methods, hybrid (ML and lexicons), and graph-based methods. Among the ML methods the commonly used classifiers are Naïve Bayes (NB), Maximum Entropy, Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Random Forest (RF), kNN (k-Nearest Neighbours) and Conditional Random Field (CRF). SA work mostly uses the supervised learning paradigm. As features, one can use n-grams, parts of speech (POS), sentence structure, word order, sentiment lexicons, or emoticons. Data are usually preprocessed. This may include tokenisation, removal of Twitter tags, POS-tagging, capturing negation, n-gram generation, statistical treatment, term frequency–inverse document frequency, stemming and feature space reduction. As measures (see III.C for definitions) the authors report mostly the accuracy or the F1 measure. However, if the classes are not of similar size, one should use the Matthews correlation coefficient [7]. Tab. I

TABLE I. TWITTER SA METHODS QUALITY OVERVIEW.
F1 and Acc (accuracy) are defined in Section III.C.

Source, Year	Method	Quality
[13], 2009	NB	Acc =0.827
[23], 2010	MNB	Acc =0.64
[5], 2010	SVM	Acc =0.813
[9], 2010	kNN	F1=0.86
[4], 2012	SVM	Acc =0.881
[32], 2012	NBSVM	**Acc =0.93**
[2], 2014	Perceptron	F1=0.78, Acc =0.77
[22], 2017, 3-class task A	CRF	F1=0.54, Acc =0.62

TABLE II. TEST DATASET OVERVIEW.
Column All also shows language split as % of all data (out of 963) and Column Negative shows % with respect to language Totals.

	About Worker	About Company	All	Negative only
comments	541	422	963	253
unique comments	507	354	854	225
tokens	7291	3758	11039	4163
unique tokens	2248	1391	3063	1617
average tokens per comment	14.4	10.6	12.9	18.5
max tokens per comment	119	103	119	119
median tokens per comment	8	7	7	10
German	201	143	344 (41%)	129 (32%)
French	173	145	313 (34%)	96 (28%)
English	133	66	195 (24%)	28 (12%)

summarises a selection of results achieved using ML on Twitter data and shows that accuracies of up to 93% are realistic. A careful study of the literature reveals various trade-offs in terms of feature and algorithm engineering. One can tune based on the size and diversity of the feature space, select the best ML method, or provide more training data. Lexicon and grammar-based enrichment are also useful.

III. MATERIALS AND METHODS

A. Gigworker and company ratings

Tab. II gives an overview of the 963 statements in German, French and English used in our experiments. The dataset was generated via manual curation from a larger set of data provided by the gig company (circa 15'000). Gig worker and employer comments are associated with a star rating (1-4 stars) with 1-2 being poor and 3-4 good. All data were passed into the Google function DETECTLANGUAGE [14]. This showed the initial language proportions German: French: English: other of feedback on the employer being 9468:1095:413:50 and of 3818:285:95:35 for comments on the gig workers, with many marked as undefined. German was the dominating language (87%), with French second (9%), then English (3%) and some Swiss German and Italian. Our target, negative statements marked with 1 or 2 stars, were underrepresented (5.5%). We anonymised the comments using simple regular expressions (replacing names with 'Person X'), removed numeric information, lowercased all and then removed empty comments. We saw that wrong language assignment had two reasons: the comment was very short or the vocabulary was used in several languages, e.g. *super service* (English or German) or *tip top* (German, French). We corrected the language assignment errors manually. We did not remove duplicated text as it came from independent workers and companies and reflects the business scenario.

Sentiment annotation was performed collaboratively by three team members with the goal of reaching circa 1000 comments in proportions roughly as shown in Tab. 2. Annotation stopped as soon as we had enough comments or slightly more. After first tentative annotation and discussions with the company, the team member with domain experience adjusted the annotation so that information with negative shading, which is potentially of use to the company, was marked as *negative* while the remaining statements (positive or neutral) as *other*. Where the sentiment was not clear, the team reached a joint decision. Typical negative comments in German were *Kommunikationsmangel* (poor communication), *nicht am Einsatz erschienen* (no show) and as other we classified *effiziente Arbeit in kleinem Team* (efficient work in a small team). Positive English comments were *great place!!* and negative *The management of my team was very bad. Very different from the other teams were (sic) it looked much more organized and tasks were clear and management was organized.* In French the workers would write *un bilan très positif, une bonne atmosphère, qui a rendu cette mission encore plus agréable* (very good overall, good atmosphere which made this job even more pleasant) or *mauvaise organization mais personnel très convival* (poor organisation but very pleasant staff). We see a variety of comment styles (long and short) and deviations from grammar and spelling rules. Out of 963, 253 comments are annotated as negative (26%) and 710 as other. Employers write longer comments and use a wider vocabulary range than the workers. Comments marked as negative are longer as they often contain a positive and a negative statement or an explanation of what went wrong. We did not detect any comments containing irony or sarcasm, but saw some figurative/informal language including German *hammer* (excellent) and French *heleine (sic) de hyene crevé* (bad breath).

A quick vocabulary frequency analysis for the main language, German, showed 13 out of 20 words were shared between the workers and companies in the most common 20 words in both groups. German examples are *dank, Einsatz, erschienen, freundlich, gerne, gut, nicht, sehr, super, Team, wieder* (thanks, job, turned up, friendly, willingly, good, not, very, great, team, again). As the available dataset is small, we cannot really do a meaningful cross-language comparison or worker/company comparison.

We used scikit-learn *tfidfvectorizer, ngram range(1,3)*, which removes punctuation, turns text to lowercase and produces word n-grams of up to three words. Stemming initially used the *NLTK* snowball stemmer [16] but was later abandoned, as it had a bad impact on the prediction accuracy (1 to 4% lower when we used stemming). We compared the star ratings with the manual annotation. We classified 1-2 stars as negative and 3-4 stars as other. It turned out, however, that the Pearson correlation between the comment and the star rating was low, around 0.65. We conclude that it is not possible to use the star rating as a proxy for sentiment. There could be three reasons for the low correlation. One is that the rating refers to the overall assessment and the comment contains additional information which only refers to some negative aspects of the job. Alternatively, the users were confused, as we know from the gig company that the users gave feedback that they did not understand what they were meant to write in the comment. Finally, in Switzerland one is rather cautious with expressing praise and 1-2 stars can be interpreted as good performance.

B. Twitter Data for the Baseline

Twitter data were used to train the model for gig sentiment prediction. We used 27’659 sentiment annotated tweets for three languages: 7369 German from the SB10k corpus [8], 4290 French from the *Canéphore* corpus [17] and 16’000 English from the Sentiment 140 corpus [13]. The tweets had 9979 negative (36%) and 17’680 other lines. Tweets were cleaned of numbers, URLs, @username and #tags and further processed the same way as the gig data.

C. Methods

Switzerland has four official languages and a large proportion of foreign nationals and migrant workers, beside the four official languages. English is used very frequently in communication. Our approach can cope with the main spoken languages, German and French, and English on top of that. In tests leading up to the results we present here, we first tested a language-specific approach for English, German and French, with stemming and without stemming. We saw only very insignificant performance differences between the multilingual and single-language approaches and therefore chose the multilingual approach without stemming, as language separation was slow and unreliable and cannot be used easily in an industrial scenario. We tested three approaches to sentiment assignment on gig data: as baseline - machine learning (ML) with Twitter training data and gig data as test data, a hybrid approach [18] using the Semantria for Excel plugin, and ML with 10-fold cross validation on hand annotated gig data. Additionally for the two best ML methods, we also used a 3-fold validation. We used Python 3.6.1 with scikit-learn version 0.19.1 [31][3]. The tests were carried out on a Lenovo G50 running Windows 10 and took a negligible amount of time. We pre-processed the data minimally, only to guarantee worker anonymity and remove digits (time, telephone number, etc.). The features were generated by the scikit-learn *tfidfvectoriser (n-gram size 1-3)*.

Tfidf [20] intends to capture the relative importance of a term (or in this case of an n-gram) in a document collection. The term frequency component *Tf(t,d)* in the simplest version is the raw count of a term in a document, i.e. the number of times that term *t* occurs in a document *d*. Often this count is scaled or normalised, for instance by dividing by the total number of terms in a document or logarithmically. The inverse document frequency component *Idf(t,D)* [20] shows the relative importance of a term in a document collection *D*. In its simplest version *idf* is calculated as

$$idf(t,D) = log \frac{N}{|\{d \in D : t \in d\}|} \tag{1}$$

where *N* is the count of documents in a corpus, i.e. *N* = |*D*|. *N* is divided by the number of documents where the term appears. Then *tfidf* [20] is a product of *tf* and *idf*:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D). \tag{2}$$

Further measures use the concept of true positives (*TP*), true negatives (*TN*), false positives (*FP*) and false negatives (*FN*). Accuracy (*Acc*) is defined as

$$Acc = \frac{TN+TP}{TN+TP+FN+FP} \tag{3}$$

and the Matthews correlation coefficient (*MCC*) which is a balanced measure where classes are of unequal sizes, can be calculated as

$$MCC = \frac{TP*TN+FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{4}$$

The *F1* score (harmonic mean of precision and recall) is

$$F1 = \frac{2TP}{2TP+FN+FP}. \tag{5}$$

F1 does not consider *TN*, and is therefore not ideal in our scenario of imbalanced data.

In our Twitter experiments, similar to the reviewed literature, we report the accuracy (*Acc*), the Matthews correlation coefficient (*MCC*) [21][7] and the confusion matrix. For gig data we also report the *F1* score. We use 10-fold cross validation with gig data and also, to see if overfitting does not take place, run 3-fold validation on the two best ML-methods (SVM and RL). In cross-validation the data are split into 10 folds (or 3 folds) and in each run 1/10th (or 1/3rd) is used as test and the rest as training data. We return an average of the results of 10 runs (or 3 runs). The pipeline has two parts: *tfidf* generation and classifier invocation with the training and with the test set. We show the confusion matrix with *TN/FN* in row 0 (our target class) and *TP/FP* in row 1, as $\begin{bmatrix} TN & FN \\ FP & TP \end{bmatrix}$. Our target class negative corresponds to true negatives and class other to positives. The perfect outcome on the gig data (963 lines) with 253 negatives and 710 others would be the matrix $\begin{bmatrix} 253 & 0 \\ 0 & 710 \end{bmatrix}$.

D. Machine learning (ML)

We use the following methods from scikit-learn with default parameters, showing in brackets the mandatory parameters:

- LSVC, the support vector classifier (SVC) with a linear kernel (C=1)
- RSVC, SVC with the Radial Basis Function kernel (C=1, gamma=0.1)
- MNB, Multinomial NB
- KNN, K-Nearest Neighbours Classifier
- Tree, decision tree classifier
- LR, linear Logistic Regression (C=1e5)

Table III. Classification with Twitter as training set showing Acc, MCC and the confusion matrix (CM). As last line we show Semantria results.

Method	Acc	MCC	CM
LSVC	**0.775**	**0.321**	$\begin{bmatrix}66 & 187\\ 30 & 680\end{bmatrix}$
RSVC	0.731	0.069	$\begin{bmatrix}15 & 238\\ 21 & 689\end{bmatrix}$
MNB	0.740	0.089	$\begin{bmatrix}6 & 247\\ 3 & 707\end{bmatrix}$
KNN	0.684	-0.032	$\begin{bmatrix}20 & 233\\ 71 & 639\end{bmatrix}$
Tree	0.724	0.239	$\begin{bmatrix}94 & 159\\ 107 & 603\end{bmatrix}$
LR	**0.759**	**0.268**	$\begin{bmatrix}63 & 190\\ 42 & 668\end{bmatrix}$
Semantria	**0.8**	**0.42**	$\begin{bmatrix}78 & 17\\ 175 & 693\end{bmatrix}$

- RF, Random Forest classifier, an ensemble method, was only used with the gig data (n_jobs=2).

IV. RESULTS

A. *Twitter Baseline*

Tab. III shows that Twitter as a training set is not appropriate. The maximum accuracy is around 0.775 and the Matthews correlation coefficient is disappointingly low, 0.321. The target class *negative*, *TN* (top left in each matrix) is not reliably classified as negative while the outcome for the class *other* is better (*TP*, bottom right of each matrix).

B. *Semantria*

Semantria [18] uses a hybrid approach: a mixture of dictionary-based category assignment and machine learning [27][29]. In the gig data it performed slightly better than the Twitter baseline, see Tab. III bottom entry. We saw 51 dictionary terms reported as the basis of polarity assignment. Example negative English terms were: confusing, dirty, disorganisation, emergency, ill. We could test 1000 items for free. The disadvantage was that it was necessary to split the data according to language, perform the analysis for each language separately and then merge the results. Incidentally, Semantria added its own language labels which were sometimes incorrect.

C. *Experiment with the gig dataset using cross validation*

Tab. IV summarises the results, with top accuracy around 0.869 and MCC around 0.646 using 10-fold validation, with logistic regression and linear SVM performing similarly. The F1 measure is around 0.91. We also show 3-fold cross-validation which produces poorer results, as we are training on a smaller data set. As we aim to find negative statements very accurately, further research is needed to deliver a method of higher reliability. Ideally, the accuracy should be around 0.95. Similarly to the Twitter baseline, we observe that the class *negative* is harder to predict accurately than the class *other*.

Table IV: Classification with gig data showing Acc (accuracy), MCC, F1 and the confusion matrix (CM).

	10-fold validation				3-fold validation			
Method	Acc	MCC	F1	CM	Acc	MCC	F1	CM
LSVC	**0.866**	**0.635**	**0.913**	$\begin{bmatrix}156 & 97\\ 32 & 678\end{bmatrix}$	**0.854**	**0.597**	**0.906**	$\begin{bmatrix}143 & 110\\ 31 & 679\end{bmatrix}$
RSVC	0.742	0.116	0.851	$\begin{bmatrix}6 & 267\\ 1 & 709\end{bmatrix}$				
MNB	0.773	0.318	0.866	$\begin{bmatrix}35 & 218\\ 1 & 709\end{bmatrix}$				
KNN	0.833	0.543	0.891	$\begin{bmatrix}146 & 107\\ 54 & 656\end{bmatrix}$				
Tree	0.829	0.540	0.888	$\begin{bmatrix}151 & 102\\ 62 & 648\end{bmatrix}$				
LR	**0.869**	**0.646**	**0.914**	$\begin{bmatrix}166 & 87\\ 39 & 671\end{bmatrix}$	**0.862**	**0.623**	**0.910**	$\begin{bmatrix}156 & 97\\ 36 & 674\end{bmatrix}$
RF	0.839	0.552	0.897	$\begin{bmatrix}131 & 122\\ 33 & 677\end{bmatrix}$				

V. Discussion

We tested three approaches: ML with training data from Twitter, a hybrid approach, and ML with 10- and 3-fold cross validation on gig data. The hybrid approach, Semantria, was similar in performance to training on Twitter (both had an accuracy <= 80% with a low MCC <= 42%). Training the classifiers on our manually annotated data improves performance. We now have an accuracy of nearly 87% with MCC of almost 65% and F1 of 91%. When comparing the ML methods we see Linear SVN and logistic regression performing the best in both ML scenarios. For completeness, we also carried out an experiment mixing Twitter and gig data as training data, which was not beneficial (details not shown). Also the effort of separating languages does not pay off, which was visible in the experiments leading to the result we report (not shown). Using stemming, which is also language specific, was not helpful either.

VI. Future Work

There are several open avenues for this research. The first one will be to perform manual annotation on another subset of data. The second option is adding words and expressions from other sources with clear negative meaning to our gig dataset to increase the n-gram range of negative vocabulary. Further options include part of speech tagging and the tuning of the feature space and ML methods themselves. We will also work on the mathematical characterisation of features allowing us to capture the vocabulary overlap between the test and training data. This will help us predict the usefulness of a training set in the classification of test data.

VII. Conclusions

To our knowledge this is the first reported multilingual sentiment analysis experiment in the gig work domain. We showed that Twitter data sets have similar linguistic characteristics to our data but the vocabulary used is too different to support reliable use of Twitter as training data. We demonstrated that manual annotation of a subset of data gives a good basis for the use of machine learning in this domain and we achieved a classification accuracy of almost 0.87 with a high Matthews coefficient of 0.65 and F1 of 0.91.

References

[1] A. Alkhatib, M.S. Bernstein, and M. Levi, “Examining Crowd Work and Gig Work Through the Historical Lens of Piecework,” Proc. 2017 CHI Conference on Human Factors in Computing Systems, ACM Press 2017, pp. 4599–4616, https://doi.org/10.1145/3025453.3025974.

[2] N. Aston, J. Liddle, and W. Hu, “Twitter Sentiment in Data Streams with Perceptron,” Journal of Computer and Communications, vol.2(03), 2014, pp. 11–16. https://doi.org/10.4236/jcc.2014.23002.

[3] J. Avila, Scikit-learn cookbook: over 80 recipes for machine learning in Python with scikit-learn, Packt Publishing, 2017.

[4] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, and V. Varma, “Mining sentiments from tweets,” Proc. 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, 2012, pp. 11–18.

[5] L. Barbosa and J. Feng, “Robust sentiment detection on twitter from biased and noisy data,” Proc. 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, pp. 36–44, 2010.

[6] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, A practical guide to sentiment analysis, Springer, 2017.

[7] D. Chicco, “Ten quick tips for machine learning in computational biology,” BioData Mining, vol. 10(1), pp. 35, 2017.

[8] M. Cieliebak, J. Deriu, D. Egger, and F. Uzdilli. “A Twitter Corpus and Benchmark Resources for German Sentiment Analysis,” Proc. Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, pp. 45–51, 2017.

[9] D. Davidov, O. Tsur, and A. Rappoport, “Enhanced sentiment learning using twitter hashtags and smileys,” Proc. 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, pp. 241–249, 2010.

[10] C. Dhaoui, C.M. Webster, and L.P. Tan, “Social media sentiment analysis: lexicon versus machine learning,” Journal of Consumer Marketing 34(6):480–488, 2017, https://doi.org/10.1108/JCM-03-2017-2141.

[11] R. Feldman, “Techniques and Applications for Sentiment Analysis,” Commun. ACM, vol. 56(4), pp. 82–89, 2013, https://doi.org/10.1145/2436256.2436274.

[12] A. Giachanou and F. Crestani, “Like it or not: A survey of twitter sentiment analysis methods,” ACM Computing Surveys (CSUR), vol. 49(2), pp. 28:1-41, 2016, http://doi.acm.org/10.1145/2938640.

[13] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” CS224N Project Report, Stanford 1(2009):12, 2009.

[14] Google, “Google Docs editors help DETECT-LANGUAGE,” 2018, https://support.google.com/docs/answer/3093278?hl=en

[15] B. Greenwood, G. Burtch, and S. Carnahan, “Unknowns of the Gig-economy,” Commun. ACM, vol. 60(7), pp. 27–29, 2017, https://doi.org/10.1145/3097349.

[16] N. Hardeniya, NLTK essentials, Packt Publishing Ltd, 2015.

[17] J. Lark, E. Morin, and S.P. Saldarriaga, “CANEPHORE: un corpus français pour la fouille d’opinion ciblee,” Poster, TALN 2015, https://hal.archives-ouvertes.fr/hal-01169293, https://github.com/ressources-tal/canephore.

[18] Lexalytics, “Sentiment extraction - measuring the emotional tone of content,” Whitepaper, consulted 27.04.2018, www.lexalytics.com.

[19] B. Liu, Sentiment analysis : mining opinions, sentiments, and emotions, Cambridge University Press, 2015.

[20] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to information retrieval, Cambridge University Press, 2008.

[21] B.W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” Biochimica et Biophysica Acta (BBA) - Protein Structure, vol. 405(2), pp. 442–451, 1975, https://doi.org/https://doi.org/10.1016/0005-2795(75)90109-9.

[22] C. Onyibe and N. Habash, “OMAM at SemEval-2017 Task 4: English Sentiment Analysis with Conditional Random Fields,” Proc. 11th International Workshop on Semantic Evaluation, pp. 670–674, 2017.

[23] A. Pak and P. Paroubek, “Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives,” Proc. 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, pp. 436–439, 2010.

[24] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs Up?: Sentiment Classification Using Machine Learning Techniques,” Proc. ACL-02 Conference on Empirical Methods in Natural Language Processing,” vol. 10, Association for Computational Linguistics, EMNLP ’02, 2002, pp. 79–86.

[25] F.A. Pozzi, E. Fersini, E. Messina, and B. Liu, Sentiment Analysis in Social Networks. Morgan Kaufmann, 2016.

[26] E. Pustulka-Hunt, R. Telesko, T. Hanne, “Gig Work Business Process Improvement,” 6th International Symposium on Computational and Business Intelligence (ISCBI 2018), Basel, Switzerland, 2018.

[27] F.N. Ribeiro, M. Araujo, P. Gonçalves, M.A. Gonçalves, and F. Benevenuto, “SentiBench – a benchmark comparison of state-of-the-practice sentiment analysis methods,” EPJ Data Science, vol. 5(1), pp. 23, 2016, https://doi.org/10.1140/epjds/s13688-016-0085-1.

[28] S. Rogers and M. Girolami, A first course in machine learning, CRC Press, 2016.

[29] J. Serrano-Guerrero, J.A. Olivas, F.P. Romero, and E. Herrera-Viedma, “Sentiment analysis: A review and comparative analysis of web services,” Information Sciences, vol. 311, pp. 18 – 38, 2015, https://doi.org/10.1016/j.ins.2015.03.040.

[30] Twitter. 2018. Twitter is what’s happening in the world and what people are talking about right now, consulted 27.04.2018, www.twitter.com.

[31] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn: Machine learning without learning the machinery," GetMobile: Mobile Computing and Communications, vol. 19(1), pp. 29-33, 2015.

[32] S.I. Wang and C.D. Manning. 2012. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In ACL (2). pages 90–9