Semi-supervised Sentiment Analysis for Under-resourced Languages with a Sentiment Lexicon

Peng Liu
Dep. of Computer Science, NTNU
7491 Trondheim, Norway
peng.liu@ntnu.no

Cristina Marco
Dep. of Computer Science, NTNU
7491 Trondheim, Norway
cristina.marco@ntnu.no

Jon Atle Gulla
Dep. of Computer Science, NTNU
7491 Trondheim, Norway
jon.atle.gulla@ntnu.no

ABSTRACT

This paper presents the results of using semi-supervised sentiment analysis on an under-resourced language such as Norwegian. To perform these experiments, two external resources have been used: an available training corpus containing Norwegian reviews from major newspaper sources (NoRec) [23], and a newly created general sentiment lexicon for Norwegian, as presented in [12]. The results of our experiments show that the performance improves significantly when the sentiment lexicon is used. Besides, the best results are obtained using Support Vector Machines (SVM) as the machine learning algorithm used for training with an AUC score of around 92%. An alternative statistical measure was used for evaluation, Area Under ROC Curve (AUC), in order to deal with the highly imbalanced nature of the dataset.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; *Content analysis and feature selection*; *Language models*;

KEYWORDS

Sentiment analysis, sentiment lexicon, under-resourced languages, feature extraction, classification

1 INTRODUCTION

Sentiment analysis, or the automatic interpretation of the positive or negative orientation of a text, is now a widely used technique for several intelligent applications.

There are two main methods to do sentiment analysis. A widely used method is to use a big training corpus to train a supervised learning algorithm. The main challenge of this approach is cross-domain sentiment analysis. As soon as the trained model is used on different corpus the performance of the analysis drops abruptly. The second method makes use of a sentiment lexicon in order to perform sentiment analysis on any type of text. Very frequently a rule-based sentiment analysis algorithm is used in this approach, which simply averages the number and/or weights of the polarity words in the text.

A common challenge of both approaches is the lack of sufficiently big and representative training corpora and sentiment lexicons. Despite the fact that the number of resources for the English language is enormous, the reality is that resources for other languages are still quite scarce. Training corpora require annotation and they are usually domain dependent. Besides, general sentiment lexicons are very expensive to build and in most languages they are not easily available.

The aim of this paper is two-fold. Firstly, we want to present the results of using semi-supervised machine learning on an available training corpus. Secondly, we seek to determine the impact of using a general sentiment lexicon for semi-supervised learning. We will perform these experiments on a low-resource Scandinavian language as Norwegian.

The contents of this paper are as follows. Related work is presented in Section 2. Then, the method presented in this paper is presented in Section 3 Our experiments and results are described in Section 4. This paper concludes with a brief discussion in Section 5.

2 RELATED WORK

A number of machine learning and lexicon-based approaches for sentiment analysis have been proposed in recent research. With respect to the first technique, most approaches use classification algorithms to determine the polarity of a text, such as Support Vector Machines (SVM), Bayesian Networks, and decision trees, among others. For example, a supervised approach was presented by Habernal et al. [8] in which they explored mutual information, information gain, chi-square, odds ratio and relevancy score. SVM has been employed to attain 73.85 % f-measure. A manually tagged Facebook dataset was employed for evaluation which may be a source of bias. Sentiment polarity categorization using information theoretic approaches was explored by Lin et al. [11]. Approaches such as information gain, chi-square were applied in completely supervised experimental settings. The sentiment scores were computed by determining the correlation of a term with positive and negative labels, respectively. Term frequency was incorporated to intensify the feature weight. An accuracy ranging from 80.65 to 82.80 % was achieved on different product review datasets. However this approach is highly domain specific, needs labeled data for training and does not handle singularities. In turn, Singh and Husain [20] evaluated three supervised machine learning algorithms namely SVM, Naive Bayes and Multi-Layer Perceptron (MLP). The best performance results of 81.15 % accuracy were achieved for SVM on a movie review dataset. Each of these supervised algorithms has its pros and cons such that nominal attributes and missing values must be processed for SVM. Besides, Naive Bayes assumes attribute independence that might not always be the case, whereas MLP needs more training data and execution time.

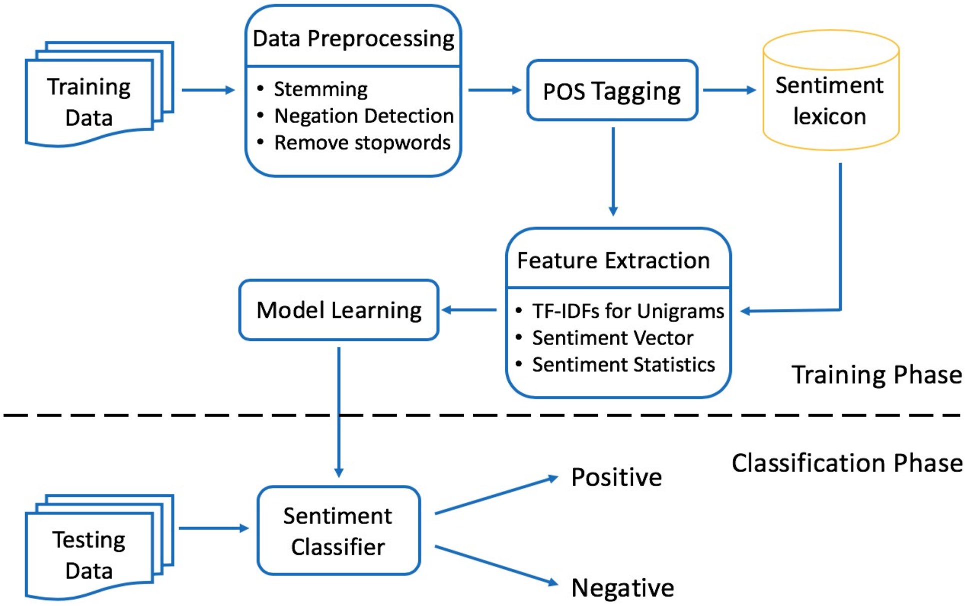

Figure 1: Overview of the proposed sentiment analysis approach.

Regarding lexicon-based techniques, SentiWordNet is one of the most generally used sentiment lexicons in the literature [14, 18]. This lexicon is based on WordNet and it contains multiple senses of a word. Besides, it provides a positive, objective, and negative value for each sense. In a semi-supervised approach, a sentiment sense inventory was built by Ortega et al. [16] based on SentiWordNet (SWN) scores. Rule-based labeling was employed to label the SWN scores into five categories. Adjectives, adverbs and verbs were utilized to achieve 50.17 % f1-score on tweets dataset. Low performance level and ignorance of nouns as semantic words are two of the major problems in this research. Ohana and Tierney [15] investigated sentiment orientation using SentiWordNet with SVM using adjectives, adverbs and verbs as candidate features in a semi-supervised manner. The feature weight was computed by considering the term position relative to the total number of terms in the document. They achieved 69.35 % accuracy on a movie review dataset. A constant value was manually adjusted to optimize the feature weight, and nouns were not included in the list of candidate features to be used as semantic words. Another semi-supervised approach was presented by Bhaskar et al. [3] in which they identified the emotions using WordNetAffect and SWN followed by SVM classification using term frequency in SVM vectors. SWN, SenticNet and a list of positive/negative words were incorporated with SVM by Chikersal et al. [4]. SentislangNet was constructed by Pandarachalil et al. [17] using SWN and SenticNet with a slangs dictionary. Ghosh and Kar [7] utilized adjectives adjacent to nouns as sentiment features based on SWN.

There are a number of lexical resources for English sentiment analysis, such as WordNet-Affect [22], SentiSense [6], Opinion Lexicon [10], Subjectivity Lexicon Riloff and Wiebe [19] and MPQA Opinion Corpus [24], etc. However, for under-resourced languages like Norwegian, it is challenging to find training corpora or sentiment lexicons. Recently, Velldal et al. [23] have released a Norwegian Review corpus which can be used for evaluating sentiment analysis algorithms. This corpus will be used in our experiments.

3 METHOD

Similar to the general frameworks of sentiment analysis, the input to our pipeline are datasets from specific data sources and the output a unique polarity score for the input document. The overview of the proposed approach is presented in Figure 1. After receiving the data from data repositories, the framework transmits the data to a data pre-processing module. After these data are part-of-speech tagged, the tagged output is conveyed into the feature extraction module, that enriches the text with sentiment information from a sentiment lexicon. Finally, the classifier assigns the input test set with sentiment polarities after the model learning process.

3.1 Data pre-processing

In order to convert the unstructured data into machine readable format, an extensive pre-processing procedures is required. Specifically, we apply the following strategies:

- Stemming and lemmatization are general means to avoid different forms of a word to appear in the same document, especially when dictionary lookup needs to be performed. Stemming usually obtains the stem of the word by removing derivational affixes. In contrast, lemmatization reduces the word to its lemma by considering the use of a vocabulary and morphological analysis of the word. Even though we are aware that lemmatization can be more effective in the early stages of data pre-processing, lemmatizers are hard to find for the Norwegian language. For this reason, NLTK was used as a stemming tool[1] to handle different word variants in Norwegian.

- Stop words are usually semantically empty, and thus they should be removed from the original documents. For Norwegian NLTK was used to perform this filtering[2].
- Negations are also crucial in handling polarity shift problem in sentiment analysis. If negation appears in a sentence, we should consider if the sentiment score needs to be reversed or not. Thus, in this paper, before the removal of the stop words, we manually keep the following negations in Norwegian: *ikke,ikkje, ei, nei, aldri, neppe, ingen, inga, intet, inkje*, that respectively mean *not, no, never, hardly, none, any.*

3.2 Part of speech tagging

Part-of-Speech (POS) tagging is the process of assigning a morphosyntactic category to each word appearing in a given text. Identifying POS tags is also a key procedure to sentiment classification tasks as it can help to distinguish different sentiment polarities with different sentiment scores.

Most POS tagging tools are designed for English language. In other languages, such as Norwegian, these tools are scarce. Fortunately, the work in [12] offers us a POS tagger for Norwegian bøkmål, which will assign each word in the corpus with a tag. The tag will then be projected to sentiment lexicon tag according to the mapping defined in Table 1. Note that we do not incorporate adverbs in our experiments because there are no adverbs in the sentiment lexicon.

Table 1: POS tags in [12] mapping to sentiment lexicon tags.

POS	POS tags in [12]	Sentiment lexicon tags
Adjective	ADJ (Adjective)	'a'
Verb	VERB (Verb), AUX (Auxiliary)	'v'
Noun	NOUN (Noun), PRON (Pronoun), PROPN(Proper Noun)	'n'

3.3 Features

Three kinds of features are extracted from the datasets by using a feature extraction module, namely TF-IDF, sentiment vector and statistical features.

Term Frequency Inverse Document Frequency (TF-IDF) is a popular and efficient scheme to determine how relevant a word is to a particular document. Intuitively, words commonly appear in a single or a small set of documents are more relevant/representative than words appears in most documents but with high term frequency. Such representative words are usually assigned with high TF-IDF score. On the contrary, commonly occurring words are always assigned a low TF-IDF score. Thus, the first feature is an input vector with the same length as the vocabulary size. Each element of the vector is set to a specific TF-IDF score if the word appears in the input document, and otherwise 0.

The second feature is a vector with the same length as the first feature (SV). In contrast, each element of the vector is assigned a particular sentiment score from the sentiment lexicon according to the word's part-of-speech in the input document.

Some statistical features are also important for sentiment classification (SS). Specifically, in this paper, we make statistics on:

1) The minimum/maximum sentiment score of the input document.
2) The number of negative/positive words of the input document.
3) The sum of negative/positive score in the input document.
4) If the sum of negative score is higher than the positive score.

3.4 Algorithms

In this paper we have evaluated the results of four different machine learning algorithms that are generally used in text classification: Gaussian Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM) and Neural Networks (NN). Implementations in the freely available package scikit-learn were used for these experiments.[3]

Machine learning algorithms have been widely used for sentiment analysis and text classification. Especially, SVM, that tries to find the maximum margin to separate classes, has been considered more appropriate than generative models for sentiment classification because it can differentiate mixed sentiment better [5]. However, in [9], it is suggested that a Naive Bayes classifier might be more appropriate for small training data since SVM needs a large set of training data in order to achieve a high classification accuracy. Besides, researchers in [2] adopted NB for Norwegian political news sentiment classification and achieved comparatively good results.

The reason that we choose LR as one of our baselines lies in that similarly to NB and SVM, LR is a lightweight algorithm with relatively high computational speed. Even for some tasks, LR probably performs better than other more complicated algorithms.

Neural Networks, especially deep learning, are gaining lots of attention lately due to its superiority in terms of accuracy when trained on huge amount of data. Recent studies have already employed NN to solve large-scale unsupervised or semi-supervised sentiment classification, in which each layer of a deep neural network architecture represents features at a different level [13, 21]. However, it has not yet been utilized

[1] http://snowball.tartarus.org/algorithms/norwegian/stemmer.html
[2] https://github.com/xiamx/node-nltk-stopwords
[3] https://scikit-learn.org/stable/index.html

for sentiment identifier in the Norwegian language. Thus, in this paper, we make an initial attempt and implement a four-layer Multilayer Perceptron (MLP) with 100 units in each layer to perform the given tasks. The detailed experiments are described in Section 4.

4 EXPERIMENTS

In this section, we conduct our experiments on a real-world dataset. First, we introduce the datasets, sentiment lexicon, data preparation and evaluation metrics. Then we compare the performance of sentiment classification with different algorithms. After that, the effectiveness of various features proposed in this paper will be tested.

4.1 Data sets

Two resources were only used in these experiments: a training corpus and a sentiment lexicon.

4.1.1 Training corpus. The Norwegian Review Corpus (NoReC) is the training corpus used for these experiments.[4] This corpus was created for the purpose of training and evaluating models for document-level sentiment analysis. The dataset contains more than 35,000 full-text reviews (approx. 15 million tokens) from Norwegian news sources and covering a range of domains, including literature, movies, video games, restaurants, music and theater, in addition to product reviews across a range of categories. In this dataset, each review is labeled with a manually assigned score of 1-6, as provided by the rating of the original author and following the Norwegian newspaper review tradition [23].

4.1.2 Sentiment lexicon. The lexicon used in this approach is a newly created general sentiment lexicon for the Norwegian language. In brief, the weights from SentiWordNet [1] were automatically transferred into the Danish WordNet, and the resulting resource was translated into Norwegian. The approach used to build this resource is explained in [12]. This lexicon contains 33,224 synsets and 35,035 wordsenses with information of their positive, negative or neutral polarity. Similarly to SentiWordnet, only 20% of the senses show positive or negative polarity. The distribution of synsets per morphological category in the lexicon is shown in Figure 2.

4.2 Data preparation

In order to perform the semi-supervised learning experiments, the review corpus was randomly split in a training, that amounts to 80% of the total review texts, and a test corpus, to 20%. To evaluate the effect of the type of text in the results, we performed experiments with two different versions of the review corpus. In the first experiments we used the full dataset, where reviews with 1, 2 and 3 review points were considered negative, and reviews with 4,5 and 6 were positive. In the second experiments, a simplified version of the review corpus was used, where only reviews with 1 and 2 review points were considered negative, and reviews with 5 and 6 positive. Reviews with 3 and 4 points were excluded from the dataset in order to avoid the natural ambiguous language inherent to borderline cases in which the reviewer does not clearly express whether she has a fully positive or negative opinion on the product or service.

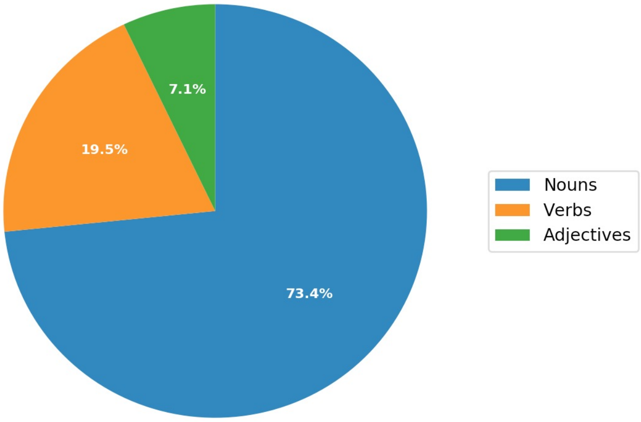

Figure 2: The distribution of synsets per morphological category in Norwegian sentiment lexicon.

Some general statistics on these two kinds of datasets can be observed in Table 2. From the disproportionate ratio, we can see the sentiment polarity distribution is remarkably imbalanced across these two datasets, which will render the standard accuracy no longer reliable. There exists many ways to alleviate such phenomena, such as up-sampling, down-sampling, change training strategy and so on. In this paper, we adopt down-sampling of our datasets with randomly removing observations from the majority class and keeping the same number of observations with the minority class. Meanwhile, instead of using the accuracy metric, we adopt Area Under ROC Curve (AUC), a well-known classification metric, to evaluate the performance.

Table 2: Some statistics of the datasets.

Datasets	Full Review Corpus	Simplified Review Corpus
#Reviews	31,671	15,713
#Pos. reviews	23,477	13,156
#Neg. reviews	8,194	2,557
Imbalance ratios	2.87	5.15

4.3 Sentiment classification results

Table 3 presents the results of sentiment classification with regard to AUC score on both datasets. Highlighted in bold is the algorithm with the best result. We can see from this table that SVM achieves the best performance followed by Logistic Regression on all datasets. Neural Networks did not show its strengths in the experiment, probably due to the

[4] https://github.com/ltgoslo/norec

limited amount of training data. We think Neural Networks are more suitable for large-scale datasets and more complex problems.

Besides, it is interesting to observe that the performances on the simplified review corpus are better than the ones on the full version. As it was mentioned before, this might be the case because in the complete dataset, reviews rated by customers with 3 and 4 review points show ambiguous opinions which actually cannot be differentiated so easily as positive or negative sentiment polarity, and thus bring extra noise to the model training process.

Table 3: The AUC score of sentiment classification results.

Datasets	Full Review Corpus	Simplified Review Corpus
NB	0.7439	0.8428
LR	0.8333	0.9257
SVM	**0.8372***	**0.9296***
NN	0.8159	0.9251

4.4 Effect of different features

In this section, we experiment on the effectiveness of classification performance with different feature combinations. The results are shown in Tables 4 and 5. We can see from both tables, the performances of SVM are superior than the other models on nearly all kinds of feature combinations on all datasets, which further verify the effectiveness of SVM model on sentiment classification.

In terms of features, similar patterns can be found on both datasets. Firstly, TF-IDF is the most important feature in our experiments because our model with TF-IDF solely achieves the best overall performance in AUC score than the model with SV or SS. Furthermore, the results deteriorate dramatically if only SV+SS are considered. On the other hand, statistical features (SS) have the lowest impact on sentiment classification for the model with TF-IDF+SV+SS improves the performance from the model with TF-IDF+SS a little but not much. Lastly, the model incorporating three input features outperforms the model with other feature combinations in AUC score suggests that all three kinds of features are still helpful in our tasks from different aspects. TF-IDF filters words with their representativeness according to TF-IDF scores in the first place. Apart from that, SV contributes to the sentiment distribution with part-of-speech appearing in sentiment lexicon. Finally, SS brings to the model useful patterns in the perspective of statistics. Therefore, our approach with TF-IDF+SV+SS presents the best AUC score in most cases in our experiments.

5 DISCUSSION AND FUTURE WORK

In this paper we investigate semi-supervised sentiment analysis using a sentiment lexicon for an under-resourced language as Norwegian. To our knowledge, this is the first paper that explores this challenge on Norwegian. The results of our experiments show that SVM perform the best. As expected, the use of features obtained from the general sentiment lexicon improves the results significantly. Interestingly, Neural Networks do not obtain competitive results. Our impression is that this might be the result of using a comparatively small dataset. We propose to use an alternative statistical measure to evaluate the performance of the machine learning algorithms, AUC, as the training corpus is highly imbalanced.

Table 4: The AUC score on full review corpus with different features.

Features	NB	LR	SVM	NN
TF-IDF	0.7346	0.8232	0.8310	0.7982
SV	0.6757	0.7365	0.7363	0.6734
SS	0.5906	0.6207	0.6223	0.6184
TF-IDF + SV	**0.7440***	0.8298	0.8348	0.8027
TF-IDF + SS	0.7356	0.8269	0.8340	0.7810
SV + SS	0.6752	0.7423	0.7428	0.6693
TF-IDF + SV + SS	0.7439	**0.8333***	**0.8372***	**0.8159***

Table 5: The AUC score on simplified review corpus with different features.

Features	NB	LR	SVM	NN
TF-IDF	0.8399	0.9176	0.9251	0.9143
SV	0.7673	0.8145	0.8147	0.7856
SS	0.6698	0.7182	0.7177	0.7237
TF-IDF + SV	**0.8438***	0.9198	0.9247	0.9176
TF-IDF + SS	0.8398	0.9229	**0.9305***	0.9093
SV + SS	0.7691	0.8299	0.7292	0.7904
TF-IDF + SV + SS	0.8428	**0.9257***	0.9296	**0.9251***

6 ACKNOWLEDGMENTS

This work was supported by the Research Council of Norway (grant number 245469).

REFERENCES

[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Vol. 10. 2200–2204.

[2] Patrik F Bakken, Terje A Bratlie, Cristina Marco, and Jon Atle Gulla. 2016. Political News Sentiment Analysis for Under-resourced Languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2989–2996.

[3] Jasmine Bhaskar, K Sruthi, and Prema Nedungadi. 2015. Hybrid approach for emotion classification of audio conversation based on text and speech mining. *Procedia Computer Science* 46 (2015), 635–643.

[4] Prerna Chikersal, Soujanya Poria, Erik Cambria, Alexander Gelbukh, and Chng Eng Siong. 2015. Modelling public sentiment in Twitter: using linguistic patterns to enhance supervised learning. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 49–65.

[5] Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *AAAI*, Vol. 6. 30.

[6] Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

[7] Monalisa Ghosh and Animesh Kar. 2013. Unsupervised linguistic approach for sentiment classification from online reviews using SentiWordNet 3.0. *International Journal of Engineering Research and Technology* 2, 9 (2013).

[8] Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2015. Reprint of "Supervised sentiment analysis in Czech social media". *Information Processing & Management* 51, 4 (2015), 532–546.

[9] Zhang Hailong, Gan Wenyan, and Jiang Bo. 2014. Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th Web Information System and Application Conference*. IEEE, 262–265.

[10] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 168–177.

[11] Yuming Lin, Jingwei Zhang, Xiaoling Wang, and Aoying Zhou. 2012. An information theoretic approach to sentiment polarity classification. In *Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality*. ACM, 35–40.

[12] Cristina Marco, Peng Liu, and Jon Atle Gulla. "Under review". Cross-lingual sentiment analysis for under-resourced languages using machine translation and sentence embeddings. ("Under review").

[13] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 572–581.

[14] Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2014. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language* 28, 1 (2014), 93–107.

[15] Bruno Ohana and Brendan Tierney. 2009. Sentiment classification of reviews using SentiWordNet. In *9th. IT & T Conference*. 8.

[16] Reynier Ortega, Adrian Fonseca, and Andres Montoyo. 2013. SSA-UO: Unsupervised Twitter Sentiment Analysis. In *Second joint conference on lexical and computational semantics (* SEM)*, Vol. 2. 501–507.

[17] Rafeeque Pandarachalil, Selvaraju Sendhilkumar, and GS Mahalakshmi. 2015. Twitter sentiment analysis for large-scale data: an unsupervised approach. *Cognitive computation* 7, 2 (2015), 254–262.

[18] Isidro Penalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia-Garcia, Miguel Angel Rodriguez-Garcia, Valentin Moreno, Anabel Fraga, and Jose Luis Sanchez-Cervantes. 2014. Feature-based opinion mining through ontologies. *Expert Systems with Applications* 41, 13 (2014), 5995–6008.

[19] Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 105–112.

[20] Pravesh Kumar Singh and Mohd Shahid Husain. 2014. Methodological study of opinion mining and sentiment analysis techniques. *International Journal on Soft Computing* 5, 1 (2014), 11.

[21] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[22] Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

[23] Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Language Resources and Evaluation*. 4186–4191.

[24] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.