Master Essay for Petter Mæhlum

This essay is about error analysis for sentiment analysis on Norwegian data. Section 1 introduces some theory behind sentiment analysis. Section 2 discusses some earlier works on SA and works using SA, both for Norwegian and other languages. Section 3 is about error analysis, both in general and for NLP, and specifically SA.

1 Sentiment Analysis

This chapter will give an introduction to sentiment analysis. It will first deal with sentiment analysis in general, and then look at the datasets that are available and relevant. Then it will look at error analysis in general, before narrowing down more specifically how error analysis is done for sentiment analysis models.

1.1 What is sentiment analysis?

Sentiment analysis (SA) is the study of how humans express their feelings or evaluations through language. These feelings are often seen as a reflection of a persons' 'private state' (Quirk et al., 1985), and are not by themselves visible; we can only infer them from how the speaker chooses their words. A person can utter the words "I am happy", and we can (hopefully) infer that the persons private state indeed is happy, but we cannot really know, and we might have to rely on other clues in order to know better. In the NLP context, these reflections are usually studied as they appear in text. However, the field of SA is far from exclusive to NLP; it is explored in other fields such as psychology and linguistics. For example appraisal theory from systemic functional linguistics, the branch of linguistics which in some ways can be said to lie closest to NLP-oriented sentiment analysis (Wiebe et al., 2005), one can talk about *polarity* and *appraisal* or *evaluation* (Thompson, 2004). In this essay I will be considering the NLP side of the problem, but research from other fields is also utilized indirectly, or directly where relevant.

1.2 Utilizing sentiment information

Understanding sentiment can help us solve many tasks. For example, we can gather information about a product and their producers, or we can summarize reviews for a ferry company. Questions like 'What part of the product do the customers like the most?' or 'What do customers feel about the resolution of this new screen?' are but a few of the many possible questions research into SA might hope to answer. Other uses for sentiment analysis include analysis of political opinion (Hammer et al., 2014) and suicide prevention (Read et al., 2012).

Imagine seeing the following sentence, taken from an album review:

(1) *Samtidig* *virker* *musikken* *gjennomtenkt* *og* *detaljert..*
At-the-same-time seem music.the thought-through and detailed.
'At the same time, the music seems well thought through and detailed'

In sentence 1, it is not hard for us to identify positive evaluative expressions, such as *gjennomtenkt* and *detaljert*. However, it is not always easy, even for experienced annotators. Bai et al. (2014) notes that even if ten people were to annotate a number of documents with sentiment-related data, the ten answers are 'almost guaranteed' to not be the same. This is perhaps a slight exaggeration, but inter annotator agreement scores from some work on SA indicate that there are indeed difficulties when doing annotation for SA (Mæhlum et al., 2019; Øvrelid et al., 2019).

1.3 Sources of Sentiment

Sentiment can be found in any text where someone expresses themselves, directly or indirectly. Sentiment can belong to artificial characters, as in prose writing, or it can be a real expression of a text's author's private states. There are, however, certain genres of writing that stand out in terms of richness of sentiment. Reviews are rich in opinions, with little irrelevant information (Liu, 2015), and several articles and datasets have been been dedicated to them (Velldal et al., 2018; Hu and Liu, 2004; Toprak et al., 2010). These reviews can be both professionally authored, such as reviews in magazines or newspapers, or more informal, as the ones found on websites such as the international movie database (IMDB). Another popular source of evaluation is the social media channel Twitter (Liu, 2015). This social media platform allows users to write their opinions about a large variety of topics, and provide valuable and relatively unfiltered language.

1.4 Granularity

Sentiment can be analysed at several levels. One can look at whole documents, both longer ones such as reviews, or shorter documents such as tweets. In these cases one is forced to look at overall sentiment, the concluding sentiment of the text, as a document can contain several different evaluations, but usually one conclusion. The next level is sentence-based SA. In these cases each sentence is a text is assigned various labels, the details of which can differ from paper to paper. The shortcoming of this method is the same as with annotation at document level: one sentence can contain several evaluations. This is especially frequent in sentences that compare or contrast different entities. This leads to the need for fine-grained sentiment analysis. Fine-grained can mean several things. It usually refers to a type of aspect-based SA, where the focus is on the aspects, that is, targets, of the sentiment expressions, but this type of analysis might include various levels of granularity when it comes to modifiers of various types, and many other traits that are identifiable at the token level. However, it does not mean that this level is without shortcomings. Fine-grained SA is

normally also bound to the sentence level, in the sense that although individual tokens can be annotated, it is not possible to annotate relations across sentences, i.e. inter-sentential annotation. It is also uncommon to annotate at the sub-token level, although this, too, can cause difficulties, especially when the target and the polar expression itself is expressed my means of the same token.

Number of polarity classes Another dimension which varies when it comes to SA is how course the classification should be. It is not difficult to find textbook examples of binary sentiment classification tasks, where one choses between 'positive' and 'negative' classes. In addition to this, some researchers employ a 5-class scheme, where the strength of a polar expression is taken into account, giving the classes 'very positive', 'positive', 'neutral', 'negative', and 'very negative' as options, or variations of these. Some authors, including Øvrelid et al. (2019), do not include a neutral class in addition to the class of sentences that do not contain any sentiment, but the term neutral can also be used to mean opinions or assessments that are not polar in nature, but still opinions.

1.5 Sentiment, Opinion and Evaluation

Several terms are used to represent a persons' private state. Words like emotion, evaluation, opinion and sentiment can all be seen. Liu (2015) mention that evaluation, appraisal, attitude, affect, emotion and mood are all related to opinion and evaluation, which are the two terms he mainly uses throughout his book. He goes on to describe 'sentiment analysis' and 'opinion mining' as largely synonymous, where the latter is more often encountered in academia. 'Appraisal or 'evaluation' is also the term used within functional grammar to indicate an '[...] indication of whether the speaker thinks that something (a person, thing, action, even,situation, idea, etc.) is good or bad' (Thompson, 2004). The term 'evaluation' is also employed by the authors of the SA works on the NoReC corpus that this essay is based on.

(2) *Samtidig* *virker* *musikken* *gjennomtenkt* *og* *detaljert..*
At-the-same-time seem music.the thought-through and detailed.
'At the same time, the music seems well thought through and detailed'

1.6 A formal definition of sentiment

So far the discussion of sentiment has been kept minimal. However, throughout this paper I will be following Liu (2015), who defines an evaluation as a 5-tuple. A sentiment expression has a target, a holder, a time of uttering and a polar expression.

Polarity *Polar* in this sense means 'positive' or 'negative'. Expressions such as 'wonderful!', 'I love it!' and 'It works well' indicate a positive polarity, while 'I do not like it.', 'It sucks!' and 'Terrible.' indicate negative polarity.

Make sure that Liu's definition is precise enough, and talk more about what is missing.

(3) *Jeg elsker jobben min.*
I love my job.
'I love my job'

Holder and target The holder, also called 'source', is the one whose private state is reflected by the expression. The target is the entity towards which the evaluation is directed. In the expression 'I love my cat', 'I' is the holder, and 'cat' is the target. Targets are often aspects of a higher-level entity. For example, a reviewer might have evaluation towards a boat cruise. In this case, the word 'cruise' might be the target in many cases, but there will also be related words, often hyponyms, of the main target. In the cruise case those might be dinner service, entertainment, quality of the beds, and many more.

Polar expression The polar expression itself is comprised of the words that indicate the evaluation. Identifying the polar expression is often more difficult than identifying the source or the target, and there are many confounding factors that can influence how a polar expression should be interpreted. In example 3, the word *elsker* is the polar expression. By using this word, the source ('I') expresses their positive private state in relation to the target ('job'). In cases like this it is clear, but polar expressions can change polarity depending on context, they can be written ironically, or they can rely on domain-specific knowledge, among other things. Details concerning types of polar expressions are discussed below.

1.7 Types of sentiment

Liu (2015) suggests a rough split into four types of sentiment. First, he splits polar expressions into two categories whether they are subjective or fact-implied. If categorized as X, there are two possibilities: emotional or rational. Emotional polar expressions are those evaluations indicated by strong 'emotional' words, such as 'like,love,hate' etc. Rational PE on the other hand, are expressions where the evaluation is to a larger degree based on utilitarity. 'useful,pretty' etc. Fact-implied polar expressions are, as the term implies, implied. In Mæhlum et al. (2019) and Øvrelid et al. (2019), their polar expression categories are based on Liu, but somewhat simplified. They see Liu's fact-implied personal category as an explicit type of evaluation, and class it together with emotional and rational to form the class they call 'evaluative'. Liu's fact implied non-personal is then called 'evaluative fact implied-non-personal'. I will use the terminology from these two papers when discussing the datasets, but refer to Liu's stricter hierarchy when discussing more fine-grained problems.

1.8 Specific problems with fine grained annotation

There are many possible sources of errors when it comes to fine grained annotation. First of all there is the problem of the difficulty of the task, and, in

relation to this, the already-mentioned problem of annotator consistency. Second, there are errors related to the data itself. Working with a multi-domain corpus might lead to low lexical overlap between the training set and new documents, meaning that the models will have to learn to a larger degree about patterns, and not just remember tokens it has seen during training. A third, returning problem with SA of this type is the sentence-level restraint. Even though the annotations are called token-based, the annotations are not intersentential. This leads to several problems where potential sentiment is lost or lose relevant details, because sentences that lack identifiable polar expressions are not included, even if a target is present. This is most often seen in examples such as sentence 4, in which *Det* refers back to a polar expression mentioned in the earlier sentence. This means that in any cases where a target appears in a different sentence than its related polar expression, the annotators are forced to ignore it. A different problem is that of sub-token items. Norwegian makes heavy use of compounding, also adj-noun compounds. This means that a target and its polar expression can occur in the same word. This might not have been treated equally by all annotators.

(4) *Det* *er* *den.*
That is it.
'That it is (It is like that).'

2 Sentiment Analysis in the Literature

A good and relatively recent summary of the NLP work is Liu (2015). In his introduction to SA, Liu sums up recent progress in the field, and formalizes many of the terms that are common. This does not mean that all researchers go by the definitions that are presented in his book. Some annotations schemes might simplify certain aspects of Liu's 5-tuple, while others might add to it. For example, Mæhlum et al. (2019) base their annotation efforts loosely on Liu's hierarchy, while they keep their annotations relatively simple on the sentence level.

2.1 A History of SA

SA moddeling follows NLP history in general, with statistical methods being the first and simplest models. A binary naive Bayes model can be used with a sentiment lexicon to provide a simple model. However, more recently, neural models are becoming more popular, with many different variations, including the aforementioned sentiment lexicons (Barnes et al., 2019), (Mitchell et al., 2013).

2.2 SA for Norwegian

SA for Norwegian is relatively new. Velldal et al. (2018) note that little work has been done, at least for datasets, for Norwegian, before the SANT-project.

In addition to this, some other work has been done by Norwegians on political opinion, constructing sentiment lexicons, and analysis of suicide letters (Read et al., 2012), but these works have not been on the Norwegian language itself. Other languages such as English enjoy a much better coverage not only for the application of SA tasks to interesting areas, but also when it comes to the creation of new datasets.

3 Data

In this essay I am going to look at the aforementioned NoRec$_{eval}$ subset.

3.1 The Norwegian Review Corpus

This thesis bases its experiments and error analysis on the Norwegian Review Corpus (NoReC), produced by Velldal et al. (2018). This corpus is a collection of review from three major Norwegian media houses, namely Schibsted Media Group, Aller Media and NRK. At the time of its creation the corpus contained more than 35 000 reviews, from 9 different genres, as seen in table 1. It is therefore a multi-domain corpus, which is different from other recent SA corpora. In Norway, it has been common to indicate the polarity of a review by a 6-sided die, where 1 is the worst and 6 is the best. This tradition allowed the authors of the preliminary excaminations to perform document-level experiments, further motivating later, more fine-grained experiments. The reviews included in NoReC are all professionally authored, and this is reflected in the style of writing. Most reviews have only one author, and there is mainly just one main topic per review.

3.2 NoReC eval

In addition to work done on the corpus at the document level (Bergem, 2018), a sub-corpus consisting of around 300, called NoReC$_{eval}$ texts were annotated at the sentence level (Øvrelid et al., 2019). These texts lay the foundation for an aspect-based annotation effort, which is described in the following paragraph.

3.3 NoReC-fine

NoReC$_{fine}$ (Øvrelid et al., 2019) is another annotated subset of the NoReC corpus. It is based on NoReC$_{eval}$, but with token-based annotations below sentence level. Like NoReC$_{eval}$, it is doubly annotated, and is comprised of roughly 250 documents.

4 Error Analysis

Language models and machine learning models can make mistakes. The models discussed for SA in the section above shows that although there have been

Category	
screen	13,085
music	12,410
literature	3526
products	3120
games	1765
restaurants	534
stage	530
sports	117
misc	102

Table 1: Number of reviews per category, taken from Velldal et al. (2018)

advances, especially with the arrival of neural networks, there are still errors that need to be investigated. There are several ways to look at errors. One option is to look quantitatively at the data, and inspect F_1-scores and other metrics. These metrics allow us to evaluate the models, but do not allow for much understanding about *why* the models behave the way they do. Error analysis (EA) is looking into the reasons why these errors happen.

When looking at language models or machine learning models, there are several approaches to how one can gain knowledge about the inner workings of the model itself. The ability for a model to be understood is know as its *interpretability*. Liu (2015) note that '[m]ost existing machine learning algorithms are black boxes'. The actual inner properties of the model functions are difficult to inspect. For smaller models, one possibility is to inspect the nodes of the model directly or to look at the probabilities in a strictly probabilistic model. However, this quickly becomes more difficult as the model increases in size and complexity. Some alternative ways of looking at this are *attention* and controlling the input and or examining the output of the models.

4.1 Work on Interpretability

One important agent in the fight against the unknown in NLP is the Black Box NLP workshop. The workshop aims to gather people from linguistics, neuroscience, machine learning and psychology[1], to better understand how machine learning models work. At the time of writing there have been two workshops: one in 2018 and one in 2019, held at EMNLP and ACL respectively. These workshops provide important data to try and answer questions about the inner workings of ML models.

[1] https://github.com/blackboxnlp

4.2 Attention

One way to look at the 'black box' of more advanced models is to use attentive architectures (Goldberg and Hirst, 2017). The use of attention was originally more common in image analysis, but has been tested for NLP in recent years, with researchers arguing both for and against its usefulness. Jain and Wallace (2019) find that there is little relation between attention results and actual relevance, while (Wiegreffe and Pinter, 2019) argue that attention networks can be useful in explaining models. The focus of this work will not be on attention, but on output analysis and challenge sets.

4.3 Output analysis and challenge sets

When wanting to understand a model, it is possible to look at how the model treats input, by looking at the output. Like interpolating a function based on a set of values, one can gain insight into a model by understanding the errors the model does. It is then possible to manipulate or control the input values, to see if the model produces the expected output.

4.4 Work on Error Analysis for NLP in general

Besides attention, several attempts at understanding the output of machine learning exist. There is a great spread in the types of models looked at: both lexicon based, statistic and neural methods are covered. Barnes et al. (2019) note that although recent advances in neural methods have resulted in quantitative improvements for sentiment analysis, there are still remaining challenges. Sentiment analysis is not only complex in terms of annotation, but also in terms of error analysis. It is not enough to simply look at misclassified expressions and tweak the model. In addition to the complexity of the models, there are also no clear formal framework to analyse the resulting errors. Because of the lack of a formal framework, this work will be looking at earlier attempts at EA not only for SA, but for NLP in general.

Categorizing errors One way to look at errors is to divide them into classes. McDonald and Nivre (2011) have a thorough look at the errors of dependency parsers. Part of their error analysis includes identifying which parts of speech (PoS) and dependency type (root, subject, object) the model struggles with. This allows them to identify more specifically what their models struggle with. They find that certain parts of speech are handled better by one of their models. The work mentioned above by Barnes et al. (2019) categorize the erroneous output of state-of-the art sentiment classifiers for English into 18 linguistic and paralinguistic categories.

Challenge sets Another way to look at errors is to create challenge sets. Challenge sets are collections of sentences that are known to be difficult by models to correctly classify. They can be created from naturally occurring data,

or they can be tailored specifically to adjust for certain factors to suit a specific task. Elkahky et al. (2018) look at noun-verb ambiguity, as certain present tense verb forms (excluding the third person) and the imperative of several verbs is homographous with related singular nouns, as in 'I feed the child', 'Feed the child' and 'fish feed'. They create a 30.000 sentence challenge dataset for PoS-tagging from naturally occuring cases of noun-verb ambiguity, containing many imperatives. They look at four commonly used taggers, and tested them on the challenge set. Their experiments show that the evaluation data itself can be used to evaluate models on this task more accurately, and that data targeted at this task can be used to improve by using extra training data for this task.

4.5 Work on Error Analysis for Sentiment Analysis

Despite the non-existence of a formal framework, that is not to say that there is no research on the field of EA for SA, and that suitable methods do not exist. One example is the work done by Veselovská and Hajič jr. (2013). They look at the misclassified sentences of their lexicon-based sentiment classifier. They identify several problems, and divide the errors into two large groups: system errors and errors caused by human annotators. Some of the errors they encounter include short segments, "domain-dependent evaluation", evaluative idioms not in their lexicon, emoticons and adversative constructions. They suggest creating a stop list for "items signalling non-evaluative part of the sentence. Another example is the paper by Pang et al. (2002), which mainly focuses on problem of sentiment classification itself, but also note some problems. They note that words that might seem intuitive, are not necessarily the best indicators of sentiment when used in a classifying task. They also note that what they call "thwarted expetations", where seemingly opposite sentiment is expressed before the real meanings are written, can cause problems for their models. Despite this, they had no clear methodology for analysing the various misclassifications apart from these observations.

4.5.1 Artificially Generated Challenge Sets

Gulordava and Merlo (2016) suggest another method of creating challenge sets. By permutating syntactic trees from existing treebanks, they can examine word order as a separate phenomenon. Their resulting trees are not necessarily grammatical, but allows for controlling only this one variables. Their motivation comes in part from the complexity of parsing morphologically rich languages, which tend to also have a relatively free word-order.

Challenge sets for SA Verma et al. (2018) also attempt to create a test set of about 150 sentences which are used to evaluate two models in terms of how well they treat polarity items in the two contexts of downward entailment and non-monotone quantifiers. This test set is similar to the challenge sets mentioned earlier, but artifically created using two sentence templates. In Barnes et al.

(2019), the sentences in the different error classes were gathered to form a dataset which could then be used to inspect various models.

5 Future work

This thesis will look at the misclassified output of the models for NoReC$_{fine}$, and aim to classify them according to error classes as in Barnes et al. (2019). The goal is to try and identify (if present) linguistic categories or annotation errors that might contribute to the classification difficulties, and then to look at possible causes and solutions for these problems. For future work I would like to expand on Barnes et al. (2019) and similar papers, and perform a error analysis for the new dataset presented in Øvrelid et al. (2019). I would like to look at the results from this initial error analysis to create a more specific challenge set to test various hyptheses.

References

Bai, A., Hammer, H., Yazidi, A., and Engelstad, P. E. (2014). Constructing sentiment lexicons in norwegian from a large text corpus. pages 231–237.

Barnes, J., Øvrelid, L., and Velldal, E. (2019). Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Bergem, E. A. (2018). Document-level sentiment analysis for norwegian. Master's thesis, University of Oslo.

Elkahky, A., Webster, K., Andor, D., and Pitler, E. (2018). A challenge set and methods for noun-verb ambiguity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2572, Brussels, Belgium. Association for Computational Linguistics.

Goldberg, Y. and Hirst, G. (2017). *Neural Network Methods in Natural Language Processing*. Morgan Claypool Publishers.

Gulordava, K. and Merlo, P. (2016). Multi-lingual dependency parsing evaluation: a large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4:343–356.

Hammer, H. L., Solberg, P. E., and Øvrelid, L. (2014). Sentiment classification of online political discussions: a comparison of a word-based and dependency-based method. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 90–96, Baltimore, Maryland. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. pages 168–177.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. *CoRR*, abs/1902.10186.

Liu, B. (2015). *Sentiment analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge, United Kingdom.

Mæhlum, P., Barnes, J., Øvrelid, L., and Velldal, E. (2019). Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 121–130, Turku, Finland. Linköping University Electronic Press.

McDonald, R. and Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230.

Mitchell, M., Aguilar, J., Wilson, T., and Van Durme, B. (2013). Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, London.

Read, J., Velldal, E., and Øvrelid, L. (2012). Topic classification for suicidology. *Journal of Computing Science and Engineering*, 6:143–150.

Thompson, G. (2004). *Introducing Functional Grammar*. Hodder Education, London.

Toprak, , Jakob, N., and Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse.

Velldal, E., Øvrelid, L., Bergem, E. A., Stadsnes, C., Touileb, S., and Jørgensen, F. (2018). NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

Verma, R., Kim, S., and Walter, D. (2018). Syntactical analysis of the weaknesses of sentiment analyzers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1122–1127, Brussels, Belgium. Association for Computational Linguistics.

Veselovská, K. and Hajič jr., J. (2013). Why words alone are not enough: Error analysis of lexicon-based polarity classifier for Czech. In *Proceedings of the 3rd Workshop on Sentiment Analysis where AI meets Psychology*, pages 1–5, Nagoya, Japan. Asian Federation of Natural Language Processing.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39:164–210.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not Explanation. *arXiv e-prints*, page arXiv:1908.04626.

Øvrelid, L., Mæhlum, P., Barnes, J., and Velldal, E. (2019). A fine-grained sentiment dataset for norwegian.